A systematic review of therapist effects:

A critical narrative update and refinement to Baldwin and Imel's (2013) review

Robert G. Johns Michael Barkham* Stephen Kellett David Saxon

Clinical Psychology Unit

Department of Psychology

University of Sheffield

Sheffield S10 2TN, UK

* Corresponding author.
Email address: m.barkham@sheffield.ac.uk

Keywords: Therapist effects; Practice-based studies; RCTs; Multilevel modelling; Systematic review

The published version appears in *Clinical Psychology Review* (2018),
https://doi.org/10.1016/j.cpr.2018.08.004

Received 28 November, 2017; Revised version received 14 August, 2018; Accepted for publication: 16th August 2018

Abstract

Objective: To review the therapist effects literature since Baldwin and Imel's (2013) review.

Method: Systematic literature review of three databases (PsycINFO, PubMed and Web of Science) replicating Baldwin and Imel (2013) search terms. Weighted averages of therapist effects (TEs) were calculated, and a critical narrative review of included studies conducted.

Results: Twenty studies met inclusion criteria (3 RCTs; 17 practice-based) with 19 studies using multilevel modeling. TEs were found in 19 studies. The TE range for all studies was 0.2% to 29% (weighted average = 5%). For RCTs, 1%–29% (weighted average = 8.2%). For practice-based studies, 0.2–21% (weighted average = 5%). The university counseling subsample yielded a lower TE (2.4%) than in other groupings (i.e., primary care, mixed clinical settings, and specialist/focused settings). Therapist sample sizes remained lower than recommended, and few studies appeared to be designed specifically as TE studies as opposed to maximising on the availability of large routine patient datasets.

Conclusions: Therapist effects are a robust phenomenon although considerable heterogeneity exists across studies. Patient severity appeared related to TE size. TEs from RCTs were highly variable. Using an overall therapist effects statistic may lack precision, and TEs might be better reported separately for specific clinical settings.

Introduction

Psychotherapy research has traditionally focussed on either the treatment modality or the patient when investigating the effectiveness of psychological therapies (Wampold & Imel 2015). However, most psychotherapy outcome studies employ multiple therapists that treat a range of patients, and this hierarchical structure of patients nested within therapists creates the opportunity to study the relative impact of therapists on outcomes (Wampold 2001). A number of studies have recognised the nested structure in the analysis and have shown that therapists do play a significant role in patient outcomes – a phenomenon termed as a therapist effect (e.g., Baldwin & Imel 2013; Barkham, Lutz, Lambert, & Saxon 2017; Lutz & Barkham 2015).

A therapist effect measures the similarity between the outcomes of patients treated by the same therapist and is akin to the intraclass correlation coefficient (ICC; Raudenbush & Bryk 2002). The ICC can also be interpreted as the proportion of the total outcome variance attributable to the variability between therapists, with larger ICCs reflecting greater variability. For example, an ICC of 0.05 – that is, a therapist effect of 5% – means that 5% of the variance in patients' outcomes is accounted for by the variability between therapists.

Therapist effects have been reported regardless of context or methodology, or whether the study has high internal validity as in the case of a clinical trial (e.g., Kim, Wampold, & Bolt 2006) or high external validity as in the case of practice-based (i.e., naturalistic) studies (e.g., Saxon & Barkham 2012). Evidence regarding therapist effects is important because it: (1) redresses the over-attention paid to comparing 'brands' of therapy (e.g., Barkham et al. 2017); (2) can identify the more and less effective therapists, which enables potentially better matching of patients to therapists (e.g., Boswell, Kraus, Constantino, Bugatti, & Castonguay 2017); (3) has the potential for advancing theory-practice links by identifying the characteristics and practices of more and less effective therapists (e.g., Wampold, Baldwin, Grosse Holtforth, & Imel 2017); and (4) can generate research questions for potential intervention studies aimed at reducing variability between therapists in an effort to improve overall service performance (e.g., Saxon, Firth, & Barkham 2017).

This review focusses on studies of therapist effects and seeks to review and critique the therapist effects evidence base published since the field was previously reviewed and summarized by Baldwin and Imel (2013) in their chapter in the 6th Edition of Bergin and Garfield's *Handbook of Psychotherapy and Behavior Change*. It is noteworthy that their chapter replaced previous chapters in earlier editions focusing on therapist variables (e.g., Beutler et al. 2004), thereby reflecting the increased research attention on this phenomenon.

Therapist effects: A brief history

Therapist effects appear to have first been commented on by Ricks (1974). When comparing two therapists who treated emotionally disturbed adolescents, four out of 15 (27%) treated by one therapist in comparison to 11/13 (85%) treated by the other therapist went on to develop adult schizophrenia. The adolescents called the former therapist 'supershrink' due to recognizing aspects of the therapist's actions that they felt were beneficial (Ricks 1974). These actions have been summarized in terms of the therapist providing greater "effort, greater support of clients' autonomy, use of resources outside of therapy, and better relationships with clients' parents" (Najavits & Strupp 1994; p.115). The other therapist became depressed and had very little energy for the most disturbed cases (Ricks 1974). Crucially, variability between the actions of the two therapists seemed to emerge in response to the more severely disturbed adolescents.

However, the issue of variability between therapists continued to be largely ignored in psychotherapy research. A review of 33 studies by Martindale (1978) found that the majority (21; 63%) did not recognize practitioner variability and just one study (3%) treated practitioners as if drawn randomly from the population of practitioners, which would have made the results generalizable. In light of these findings, Martindale stated that researchers were inappropriately generalizing findings beyond the practitioners involved in outcome studies.

A meta-analysis by Crits-Christoph et al. (1991) marked the first summary quantitative statement of the research evidence regarding therapist effects. This study reported that across 27 different treatment groups, therapists accounted for an average of 8.6% of the outcome variance. Wampold (2001) similarly found that therapist effects accounted for approximately 8% of the outcome variance, while the effects of specific treatments hovered around zero. Okiishi, Lambert, Nielsen, and Ogles (2003) highlighted the extent of variability in outcomes achieved by different therapists sampled from a single clinic and their results were consistent with findings from multisite studies. Brown, Lambert, Jones, and Minami (2005) reported that clients seen by the most effective therapists from a variety of treatment settings achieved three times as much change as compared with those showing least change.

In 2013, Baldwin and Imel provided the most detailed summary to date of the therapist effects literature. They identified 25 studies reporting fixed effects analyses (i.e., where comparisons are restricted to the sample of therapists used in each individual study) and 46 studies using random effects (where results can be generalised to the population of therapists). For the fixed effects studies, no summary breakdown was provided for RCT and practice-based studies regarding the

percentage split or median/mean number of therapist and patients per study. Of the 46 random effects studies, 29 were efficacy studies (i.e., trials) yielding a therapist effect of approximately 3% and 17 were naturalistic or effectiveness studies yielding a therapist effect of 7%. This difference may be explained by therapist effects being suppressed in trials due to tight inclusion criteria, adherence checks, manualization, close supervision and smaller samples of therapists. The overall average therapist effect found across the studies was 5% and the most effective therapists were twice as effective on average when compared with the least effective therapists. These random effects studies yielded a total of 1218 therapists and 14,519 patients, but the median number of therapists per study was only 9 (range 2 to 581), and in only two studies did the mean number of patients per therapist exceed 30 (Cella, Stahl, Reme, & Chalder 2011; Dinger, Strack, Leichsenring, Wilmers, & Schauenburg 2008).

The evidence base for therapist effects, therefore, increasingly supports the view that some therapists facilitate better patient outcomes than others. Hence, despite policy guidance (e.g., National Institute for Care and Clinical Excellence [NICE] guidelines e.g., NICE 2009) implying homogeneity of delivery (i.e., for problem x, apply therapy y), the therapist effects phenomenon suggests that, at the point of delivery, significant heterogeneity exists between therapists. Therapist effects also appear to prevail regardless of whether the context is a clinical trial (e.g., Huppert et al. 2001) or a study of routine clinical practice (e.g., Saxon & Barkham 2012), although the size of effect varies.

Variability in size of therapist effect

In the study by Crits-Christoph et al. (1991) the therapist effect ranged from 0% to 48% across 15 studies, while in Baldwin and Imel's (2013) review therapist effects ranged from 0% to 55% across 46 random effect studies. The heterogeneity of the studies included in these reviews is cited as the cause of the variability of effects, although currently little is known about the specific effects of different factors on the size of therapist effect (Baldwin & Imel 2013; Wampold 2007).

There are four main factors that may contribute to the size of therapist effect: the statistical approach adopted, sample size, the case mix variables included in the analyses, and the clinical setting. First, different statistical approaches adopted in studying therapist effects can lead to different results (e.g., Elkin, Falconnier, Martinavich, & Mahoney 2006; Kim et al. 2006). The statistical approach recommended to investigate therapist effects is multilevel modeling (MLM), sometimes termed hierarchical linear modeling (HLM), in which the hierarchical structure in the data is recognised (Adelson & Owen 2012). This allows for the separation of the outcome variance

between the therapist level (level 2) and the patient level (level 1) and the calculation of the proportion that is at the therapist level, which is the therapist effect (Raudenbush & Bryk 2002; Wampold & Brown 2005). MLM avoids potential Type I and Type II errors arising from single level approaches (Hox 2010), such as the use of analysis of variance (e.g., Huppert et al. 2001), although this latter study was subsequently reanalyzed using MLM (see Huppert et al. 2014). And, particularly important, MLM also controls for patient variables and case mix. The Baldwin and Imel (2013) review found that not all studies used multilevel or random effects analysis.

Secondly, large sample sizes, particularly of therapists, are required to estimate statistically reliable therapist (i.e., level 2) effects (see Maas & Hox 2005; Schiefele et al. 2017). Low power resulting from small numbers of patients in traditional outcome studies (Kazdin & Bass 1989) will also lead to under-powered therapist effect studies (Crits-Christoph, Tu, & Gallop 2003; Owen, Drinane, Idigo, & Valentine 2015). The smaller the sample size at each level of the multilevel model, the greater the risk of over or under estimating the size of therapist effects (Baldwin & Imel 2013). A key recommendation for future therapist effect studies made in the Baldwin and Imel (2013) review was for researchers to acquire larger sample sizes to avoid power issues and sampling error. Another recommendation called for more studies that were designed from the outset as a therapist effect study.

The third factor is the list of case-mix variables that are included in any analyses. For example, Okiishi et al. (2006) found that controlling for patient initial severity explained a considerable amount of the variability between therapists, while random slopes for patient severity, where the relationship between patient severity and outcome varies between therapists, has been a consistent finding of therapist effect studies (Schiefele et al. 2017). The relationship between number of sessions attended and outcome has also been found to vary between therapists. Saxon et al. (2017) found a therapist effect of 2% where patients received 2 treatment sessions and 40% where they received 20 sessions or more. The case-mix variables controlled for will have an influence on the size of therapist effect.

The final factor, linked to the third, is the clinical setting from which the data was collected. Therapist effects are based on the average patient in the sample and the mean values of case-mix variables included in the analysis. It might, therefore, be anticipated that therapist effects may differ across different types of clinical settings that reflect different patient populations. In the same way that therapist effects have been found to differ as a function of patient severity within a single setting (Saxon & Barkham 2012), the same phenomenon might be present across clinical settings

that serve patients presenting with differing clusters of psychological issues and from differing social contexts.

Review questions

In light of the above considerations, the current narrative and empirical review updates and refines the review carried out by Baldwin and Imel (2013). In particular, it provides a practical and pragmatic framework for considering therapist effects according to clinical settings, which may reflect differing levels of patient severity and different presenting conditions by the patients. This framework is consistent with the original observation of the differential impact of patient severity by Ricks (1974) and observations reported by Barkham et al. (2017) that patient severity may be a key determinant in the extent to which therapist effects are present. Accordingly, the primary aim of the review is to report the individual and combined size of ICCs reported in publications or in advance on-line from clinical trials and practice-based studies in the time period 2012 to 2016 inclusive. And, in light of Baldwin and Imel's (2013) call for larger studies and studies designed specifically to investigate therapist effects, we report on the extent to which these two recommendations have been met.

Method

Identification of studies

A systematic literature search was conducted using title and abstract searches of three online databases (PsycINFO, PubMed and Web of Science) and dates within the 5-year range January 2012 to December 2016. This included early on-line publications appearing during this time period that were subsequently published in hard copy in 2017. The start date was chosen to ensure continuity from Baldwin and Imel's (2013) review and search terms were replicated: "Therapist effects" or "therapist outcome" or "differential effects of therapists" or (therapist and "intraclass correlation") or (therapist and (multilevel or "hierarchical linear modelling" or "mixed models")) or "effective therapist" or "ineffective therapist" or "therapist variance". Reference lists of retrieved studies were also examined to identify further studies that may have been missed due to limiting the search terms as above. Preferred reporting items for systematic reviews (PRISMA) procedures were adopted (see Figure. 1; Moher, Liberati, Tetzlaff, & Altman 2009). After initial identification of studies (n = 2132), duplicates were removed, and 1566 studies examined against the inclusion criteria. Full texts of the resulting 47 studies were retrieved and examined, leading to further exclusion of 26 studies, resulting in 21 studies. One of these was a meta-analysis which was also excluded, yielding 20 studies included in the review.

Study selection criteria

Studies were included if they met the following inclusion criteria: a) published in a peer-reviewed journal, b) investigated therapist effects in a clinical population, c) published in hard copy or early on-line January 2012–December 2016, d) study samples were adults, e) written in English, and f) an empirical study examining quantitative treatment outcomes in which the focus on therapist effects was a central aim of the study. This latter criterion was premised on the recommendation that therapist effect studies should be designed primarily as studies of therapist effects rather than having therapist effects as a secondary interest (Baldwin & Imel 2013). Exclusion criteria were in keeping with therapist effects recommendations (Wampold 2005) and were the reverse of the inclusion criteria or having a primary focus on process variables (e.g., alliance, adherence) or patient dropout rates.

Quality assessment

All studies were quality assessed using a modified Downs and Black (1998) checklist. Modifications were informed by statistical (Adelson & Owen 2012), power (Schiefele et al. 2017) and reporting recommendations (Baldwin & Imel 2013) for therapist effect studies. Specifically, the power question was adapted to reflect therapist effects sampling recommendations for both therapists and patients. The sample size of therapists is generally considered most important for the reliability of therapist effects (e.g. Adelson & Owen 2012). Maas and Hox (2005) recommended at least 100 therapists for unbiased estimates of effects but a sample of 50 therapists would yield acceptable effects. Schiefele et al. (2017) recommended a sample of 1200 patients which could be derived from different combinations of therapists and patients. The required number of patients per therapists is likely to be determined by the number of therapists and also the focus of the study. As such, there is no single agreed value for the number of therapists or patients. 5 = ≥100 therapists all treating ≥10 patients each; 4 = ≥100 therapists with some or all treating <10 patients or 50–99 therapists all treating ≥10 patients each; 3 = 50–99 therapists with some or all therapists treating <10 patients; 2 = 10–49 therapists all treating ≥10 patients; 1 = 10–49 therapists with some or all therapists treating <10 patients; and 0 = <10 therapists. See Appendix A for the full checklist with details of adaptations.

The first author (RGJ) rated all articles. Two independent raters (final year trainee clinical psychologists) familiar with the original Downs and Black (1998) checklist from use of it in their own research, determined reliability of the quality checklist scores. Each rater examined a different set of 20% of all studies (i.e., 4 studies) to maximise the breadth of sampling of ratings. Each set of studies comprised one RCT study and three naturalistic outcome studies, including one from each of the highest and lowest quartile and two from the middle 50% of overall quality scores as determined by the first author. The Downs and Black (1998) sample mean (SD) scores of 14 (6.39) for RCT studies

and 11.7 (4.64) for naturalistic outcome studies were used as the quality benchmarks. See Appendix B for details of rater agreement levels.

Data extraction

As noted above, the therapist effect is derived from the intraclass correlation coefficient (ICC) defined as:

where σ_t^2 represents the variance at the therapist level and σ_e^2 represents the variance at the patient level. The ICC, therefore, gives the proportion of the total outcome variance that is associated with the therapist, which is multiplied by 100 to give the therapist effect as a percentage. For each study in the review, the ICC was reported or calculated where sufficient information was provided. To calculate an overall weighted average ICC, three parameters were considered; number of patients, number of therapists, and number of patients per therapist (Schiefele et al. 2017). Mean ICCs weighted by patient were calculated by summing the individual products of each ICC and number of patients, then dividing by the total number of patients. Similar calculations were conducted to obtain mean ICCs weighted by therapist and mean ICCs weighted by number of patients per therapist.

Results

Organisation and details of included studies

The final 20 selected studies meeting the inclusion criteria comprised either randomised control trials (n = 3; 15%) or naturalistic outcome studies (n = 17; 85%). Within the naturalistic studies, we grouped the studies according to four broad clinical settings as follows. (1) University counseling centers comprised studies defined as being based in and serving university or college students. (2) Primary care settings comprised locally run services that are deemed to be the first port of call for patients experiencing psychological difficulties. They might normally take referrals directly from General Practitioners or Family Physicians. This group also included services in the UK that were set up under the UK government's Improving Access to Psychological Therapies (IAPT) services (see Clark, 2011). The IAPT service adopts a stepped-care model in which patients are first seen by a Psychological Well-being Practitioner (PWP) who delivers a low intensity intervention (e.g., psychoeducational, self-help). If no improvement is made, patients are stepped-up to receive a high-intensity intervention (e.g., cognitive-behavioral therapy) delivered by a traditional therapist. (3) Mixed clinical settings comprised studies that sampled patients from across differing types of services and were therefore, by definition, more heterogeneous. And (4) Specialist/focused settings,

which were identified as comprising more severe and enduring patients with very defined clinical presentations and interventions.

Table 1 summarises the included studies and presents information on the number of patients, therapists, mean number of patients per therapist, SD, and lowest and highest number of patients per therapist. Individual studies are noted in terms of these descriptives as follows: patients >1200 (denoted by *); therapists >100 (denoted by ††); lesser threshold >50 (denoted by †); minimum patients per therapist >10 (denoted by ‡).

In addition, Table 1 provides information on patient diagnosis, outcome measures, treatment setting, statistical analysis, results, and quality rating. Studies are grouped by type of study (RCT or naturalistic practice-based studies), with the naturalistic studies grouped according to the four broad clinical settings. Studies within each group are listed alphabetically. Studies were qualitatively reviewed according to the above categories.

The mean number of patients per study was 6157 (range 91–48,648; SD=10,695) and the median was 3929.5 (IQR = 599–6277.5). The mean number of therapists was 187 (range 3–1800; SD = 402.2) and the median was 57.5 (IQR = 33.25–161), yielding a mean number of patients per therapist of 47 (range 6–135). The most common presenting diagnosis was depression/anxiety (n = 7; 33%) and the most common outcome measure used was the Patient Health Questionnaire-9 (PHQ-9; n = 6; 29%). The majority of studies investigated a range of different therapies within the same study and were therefore termed 'mixed psychotherapy' (n = 11; 52%). Nineteen studies (95%) used a hierarchical design, and 19 studies (95%) found a significant therapist effect.

RCT studies

Details of the RCT studies providing eligible data (n = 3) are presented in Table 1. The mean (SD) number of patients, therapists and patients per therapist were as follows: patients (M = 362.3; SD = 309.9); therapists (M = 17; SD = 18.5); and patients per therapist (M = 42; SD = 49.6). The only study meeting any of the power criteria was Goldsmith, Dunn, Bentall, Lewis, and Wearden (2015) in relation to all therapists having >10 patients each.

Naturalistic studies

Overall, for the 17 studies, the mean (SD) number of patients, therapists and patients per therapist were as follows: patients (M = 7561.7; SD =11,294.2); therapists (M = 218.6; SD = 430.4); and patients per therapist (M = 52.6; SD = 40.7). For the four groupings of service context, the equivalent values were as follows: university counseling centers (n = 5), patients (M = 6027.2; SD = 4928.5),

therapists (M = 240.2; SD=208.1), and patients per therapist (M=24; SD=14.1); for primary care settings (n = 6), patients (M = 4734.8; SD = 3560.8), therapists (M = 55.3; SD = 34.4), and patients per therapist (M = 81.8; SD = 36.7); for mixed clinical settings (n = 4), patients (M = 15,803; SD = 21,929.6); therapists (M = 531; SD = 848.3); and patients per therapist (M = 190; SD = 18.5); and for specialist/focused settings (n = 2), patients (M = 147.5; SD = 62.9); therapists (M = 17.5; SD = 10.6); and patients per therapist (M = 9.0; SD = 1.4).

In terms of the product of patient and therapists, we considered a conservative calculation of Schiefele's criterion of 1200 patients by using the lowest reported number of patients per therapist rather than the mean. As shown in Table 1 (denoted by *), nine studies met this criterion with all of them, except Hayes, McAleavey, Castonguay, and Locke (2016) and Schiefele et al. (2017), having a minimum of at least 10 patients per therapist. In terms of the number of patients, therapists, and lowest number of patients per therapist, four studies met all three criteria including number of therapists >100: Goldberg, Hoyt, Nissen-Lie, Nielsen, & Wampold 2016; Goldberg et al. 2016; Nissen-Lie et al. 2016; and Saxon and Barkham (2012). Four further studies met the three criteria when number of therapists >50: Chow et al. (2015); Firth, Barkham, Kellett, and Saxon (2015), Kraus et al. (2016); and Saxon et al. (2017).

Quality ratings

All studies exceeded the quality benchmark scores (range 20–27) and were therefore included in the review. Agreement between the two independent raters and the original rater were: rater 1, κ = 0.72, and rater 2, κ = 0.66 (both $p<0.01$). See Appendix B for the full results of the quality checklist. There was no significant correlation between year of publication and quality score for either all studies (N = 20; r = 0.35, p=.13) or practice-based studies (N = 17; r = 0.21, p = .43). For the RCTs, the mean (SD) quality rating was 22.3 (2.08), 95% CI = 20.0 to 24.7). For naturalistic studies, the mean (SD) quality rating was 24.4 (2.1), 95% CI = 23.3 to 25.4. For each of the four groups, the summary statistics were as follows: university counseling centers, M = 24 (3.1), 95% CI = 21.3 to 26.7; primary care settings, M = 25.3 (1.5), 95% CI = 24.1 to 26.5; mixed settings, M = 24.5 (1.3), 95% CI = 23.2 to 25.8; and specialist/focused settings, M = 22 (other values = 0). In sum, with the exception of the two specialist/focused studies, the mean quality of naturalistic studies exceeded those of RCT studies, with the quality ratings for primary care studies obtaining the highest mean quality ratings.

Study methods and components

Table 2 presents a summary of the methods and components used in each study design. The three trials showed a range of follow-up measure time-points, practitioner groups and analytical methods.

In the naturalistic designs, 11/17 (65%) collected pre-post data as opposed to sessional and no study collected follow-up data. The focus of the main outcome measures was mixed, split between symptom measures and those measures tapping a range of presenting issues. The most frequently controlled variable was severity, in 10/17 (59%) studies. In terms of therapists, most studies described their samples differently, suggesting a wide variation in the professional backgrounds and affiliations of therapists. Only 5/17 (29%) of studies included additional variables in their investigations of therapist effects. All but three of the naturalistic studies used 2-level analysis for calculating the ICCs.

Average therapist effect size

Tables 3 and 4 show details of the ICCs reported (or calculated if the ICC was not reported) for each model and outcome measure and the mean ICCs for each study. Converted to therapist effects (i.e., percentages), effects from individual models varied from 0.2% to 29%. The average effect across all studies, weighted by number of patients was 4.9% and by number of therapists was 5.0%. When weighted for number of patients per therapist, the effect was 5.4%. This implies that across studies, approximately 5% of the variance in outcomes was attributable to the therapist.

The average effect for the 3 RCT studies was 12.9% weighted by number of patients, 17.4% weighted by number of therapists and 8.2% weighted by number of patients per therapist, giving a therapist effect between 8.2% and 17.4%. For naturalistic studies, the mean effect was 4.7% weighted by number of patients, 4.8% weighted by number of therapists and 5.0% weighted by number of patients per therapist, giving an overall therapist effect of around 5%.

The average effects for each of the four groups was also calculated. For university counseling centers: 3.1% (weighted by patient; range 0.1–15.3%), 3.6% (weighted by therapist; range 0.5–16.2%) and 2.7% (weighted by patient per therapist; range 0.4–7.3%). For primary care: 5.1% (weighted by patient; range 0.1–22.0%), 4.3% (weighted by therapist; range 0.2–19.0%) and 4.8% (weighted by patient per therapist; range 0.2–13.8%). For mixed clinical services: 5.9% (weighted by patient; range 0.2–60.4%), 6.1% (weighted by therapist; range 0.4–57.9%) and 5.9% (weighted by patient per therapist; range 1.1–26.1%). And for specialist/focused services: 13.0% (weighted by patient; range 11.7–13.8%), 12.5% (weighted by therapist; range 10.5–14.6%) and 15.0% (weighted by patient per therapist; range 8.9–25.3%).

Therapist effect and sample sizes

To determine the association between therapist effect and the N of patients, therapists, and patients per therapist in each study, we calculated Spearman's rho for various groupings on the studies. For all studies (n = 20), correlations were significant for the N of patients (−0.63, p<.003) and for N of therapists (−0.67, p<.001). That is, the larger the N values, the lower (more conservative) the therapist effect. The N for patients per therapist was not significant (p>.05). For naturalistic studies (n = 17), only the N of therapists was significant (−0.64, p < .005). Neither the N of patients nor N of patients per therapist was significant (p values = .054 and 0.69 respectively). When only the 14 naturalistic studies using MLM were considered, both the N of therapists and N of patients were significant (−0.71, p<.004, and−0.61, p<.02 respectively).

Reporting bias

In order to assess the presence of reporting bias, a funnel plot of ICC scores against number of patients per therapist was constructed (see Figure 2). Each dot on the plot represents one of the ICCs in Tables 2 and 3 and patients per therapist was chosen as the most representative measure of sample size. Although asymmetrical due to not being able to have an ICC below zero, the graph indicates possible over reporting of large effects where the samples are small.

Review of randomised control trials

In this section, we focus on reviewing the RCTs identified in the search and on issues impacting on the heterogeneity within settings (i.e., design and aims, outcomes, variables) and possible reasons for different effects. Three studies investigated therapist effects within RCTs (Erickson, Tonigan, & Winhusen 2012; Goldsmith et al. 2015; Moyers, Houck, Rice, Longabaugh, & Miller 2016) and therapist effects ranged from 1 to 29%.

In Goldsmith et al.'s (2015) study, outcome was level of fatigue and physical functioning, and patients were randomised both to one of the three nurses and one of two treatment arms (pragmatic rehabilitation or supportive listening) with the nurses delivering both interventions. The analyses employed regression models rather than MLM and found no therapist effects in either treatment arm. Whilst it could be argued that randomisation nullified any therapist effect, the use of only three nurses made it the smallest sample in this review and the only study employing nurses. Additionally, unlike all other studies included in the review, outcome measures were more related to physical symptoms. Hence, the study appears substantially different on key design factors that make it unrepresentative of other studies in the area.

Erickson et al. (2012) also used randomisation to therapist when investigating therapist effects in pregnant substance users. Taken from a larger RCT, participants were all randomised to either manualised motivational enhancement therapy (MET) or treatment as usual (TAU). Outcomes were self-reported substance use and urine analysis and MLM found a therapist effect of 29% for the MET condition, which disappeared when one of the 10 therapists was excluded. Limitations of the study included low therapist numbers and the issue that some patients were receiving other treatments concurrently.

Moyers et al. (2016) investigated therapist effects and therapist empathy in an RCT of behavioral treatment during an alcohol reduction program. Results showed that 11% of outcome variance (i.e., alcoholic drinks per week) was associated with therapists. Empathy levels were not found to vary between therapists, but within-therapist variations were apparent across therapy sessions (e.g., during sessions of higher empathy, larger decreases in drinking behaviours occurred). A major limitation of the study was that empathy was rated by observers rather than by patients.

Review of naturalistic practice-based studies

In this section, we focus on reviewing the 17 practice-based studies identified in the search and focus on issues impacting on the heterogeneity within settings (i.e., design and aims, outcomes, variables) and possible reasons for different effects. We considered these studies in the four groupings identified earlier.

University counseling centers.

Six studies analysed data from US university counseling centres (Goldberg, Hoyt, et al. 2016; Goldberg, Rousmaniere, et al. 2016; Hayes et al. 2016; Hayes, Owen, & Bieschke 2015; Nissen-Lie et al. 2016; Owen, Adelson, Budge, Kopta, & Reese 2016). Therapist effects ranged from 0.4–19.1% with a weighted average of 2.4%. This smaller than average effect may reflect the fact that the sample comprised patients (i.e., students) who presented with less severe symptoms.

Owen et al. (2016) calculated therapist effects from three subscales of the BHM-20. Results showed therapist effects of <1% for wellbeing, 4.6% for symptom distress and 7.5% for life functioning. Although the overall therapist effect of 2.4% was relatively small compared to effects in other settings, these findings are consistent with the evidence that the more complex the presenting problems, (i.e., symptom distress and life functioning compared to wellbeing), the greater the variability between therapists. One limitation of the study was validity of the subscales with the wellbeing subscale comprising only three items and the life functioning subscale comprising four

items. Accordingly, these scales may miss both broader and more specific aspects of patient change (and thus therapist variability).

Therapist effects over time.

Two studies within this group investigated the extent to which the effectiveness of therapists varied over time (Goldberg, Hoyt, et al. 2016; Goldberg, Rousmaniere, et al. 2016). Therapist effects ranged from 0.09–1.1%. In Goldberg, Hoyt, et al.'s (2016) study, the highest and lowest 10% of therapists were classified into high performing or low performing groups. Results showed a small overall therapist effect of 0.089%, alongside an increasing discrepancy between high and low performing groups as treatment duration increased This implies that the therapist becomes more important as a function of the duration of therapy, which may be a proxy for the severity and complexity of the presenting problems.

Goldberg, Rousmaniere, et al. (2016) investigated whether effect sizes increased as therapist experience increased. MLM showed a therapist effect of 1% with effect sizes of therapists decreasing very slightly over time, with wide variation in different therapists' trajectories. Limitations included the heterogeneity of the therapists in terms of experience and treatment approach and the lack of recording of training and supervision received. Although both Goldberg studies calculated an overall therapist effect, they did not consider whether this overall therapist effect within the sample varied at different time points.

Racial diversity.

A further two studies investigated therapist effects in populations comprising a diversity of White and racial/ethnic minority (REM) clients. Hayes et al. (2015), using MLM analysis, found that the variability in therapists' outcomes was a partial function of the REM status of the patients. Two limitations of the study were the small number of therapists and patients and the single treatment centre. Hayes et al. (2016) extended the previous study across 45 university counseling services, finding a therapist effect of 3.9%. Overall, both REM and non-REM patients experienced similar levels of symptom reduction. However, the study identified some therapists as having better outcomes with REM patients than non-REM patients, while this was reversed for other therapists.

Primary care settings (including Improving Access to Psychological Therapies studies).

Six studies investigated therapist effects in UK primary care settings: IAPT high intensity (n=1; Saxon et al. 2017), IAPT low intensity (n=3; Ali et al. 2014; Firth et al. 2015; Green, Barkham, Kellett, &

Saxon 2014), and mixed (n = 1; pre-IAPT, Saxon & Barkham 2012; n = 1; IAPT, Pereira, Barkham, Kellett, & Saxon 2017). Therapist effects in these studies ranged from 0.9–9.7%.

Low intensity

Ali et al. (2014) investigated the effects of treatment characteristics by examining therapist effects in brief low-intensity psychological interventions provided by Psychological Wellbeing Practitioners (PWPs). Routinely collected outcome measures for depression and anxiety were analysed in an IAPT service. They used a three-level hierarchical structure with sessions at level 1, patients at level 2, and PWPs at level 3. Results showed therapist effects of 1% for the depression measure (PHQ-9) and 0.9% for anxiety (GAD-7). All PWPs had outcomes that were not statistically different from the 'average' PWP in the sample. These relatively low therapist effects may be attributable to the low initial severity of patients (i.e., mild-to-moderate depression/anxiety) and/or case complexity of the sample. However, the authors used a three-level hierarchical model with sessions at the lowest level and did not control for initial severity, which may have constrained the overall therapist effect and result in a more likely explanation of the low therapist effect. The lowest reported number of patients per therapist was one. Firth et al. (2015) investigated therapist effects and efficiency in PWPs in a similar IAPT service to Ali et al. (2014) and using outcome measures for anxiety, depression and functional impairment. A therapist effect of 6–7% was moderated by initial symptom severity, duration of treatment and non-completion of treatment. The most effective PWPs were found to achieve nearly twice the change per session in comparison to less effective peers. The study found a much larger therapist effects than Ali et al. (2014) in a very similar service with identical outcome measures but used a 2-level model.

Green et al. (2014) also investigated PWPs across 6 IAPT clinics and found therapist effects of 9–11% controlling for pre-treatment severity. The study by Green et al. (2014) also highlighted that therapist resilience, organisation, knowledge and confidence were associated with more effective therapists.

High intensity

Saxon et al. (2017) investigated therapist effects in a large naturalistic dataset of patients receiving counseling or CBT in an IAPT service. After controlling for case mix, a therapist effect of 5.8% was found. Completion of therapy and higher number of sessions attended were both associated with larger therapist effects and more effective therapists were found to have recovery rates twice that of the less effective therapists. There was no significant difference in the effect size between CBT and counseling.

Low and high-intensity

Saxon and Barkham (2012) used MLM to investigate therapist effects in patients receiving psychological therapy or counseling in a primary health care setting across an 8-year period (immediately before the establishment of the IAPT initiative but comprising patients who would have been very similar to those later referred to the IAPT service). Results showed a therapist effect of 6.6%. Greater initial patient severity and higher therapist caseload risk levels were associated with poorer outcomes, with the effect ranging from 1% to 10% as severity varied and the effect being reduced from 7.8% by the inclusion of therapist caseload risk. However, the least effective therapists had almost half the recovery rate of the above average therapists.

Pereira et al. (2017) analysed data from a single IAPT service, measuring patient depression outcomes and therapist self-reports of resilience and mindfulness. An overall therapist effect of 6.7% was found across high and low-intensity IAPT therapists, with more effective therapists having higher levels of mindfulness, along with resilience and mindfulness combined.

Mixed clinical settings.

Four studies compared therapist effects in mixed clinical settings – that is, studies pooling patients from a variety of settings: Chow et al. (2015), Kraus et al. (2016;), Nissen-Lie et al. (2016), and Schiefele et al. 2017). Therapist effects ranged from 1.3% to 12.9%.

Chow et al.'s (2015) main sample, derived from 45 organizations, yielded a 5.1% therapist effect using a 3-level model. They investigated a subsample of 17 therapists with 1632 patients. For this subsample, not reported in Table 1, the mean therapist caseload was 94.24 (SD = 97.40; Mdn = 46.00; minimum = 10, maximum = 335).They found that the characteristic that best predicted effectiveness was the amount of time dedicated by therapists to improving their therapeutic skills, termed deliberate practice, supporting the view that more dynamic qualities of therapists may be related to therapist effects.

Kraus et al. (2016) investigated therapist effects across a range of sub-domains of the Therapy Outcomes Package (TOP; Kraus, Seligman, & Jordan 2005) and across a wide range of treatment settings. Scores were risk-adjusted by intake score, risk score, and then with a full random forest model. The TOP yields 12 subscales and therapist effects across these outcome domains when fully risk-adjusted ranged from 1.6–18.7%, with an overall effect of 12.9%. We considered only the overall effect because of concerns in generating 12 TEs from a single study. Similar to Owen et al. (2016), the quality of life measure produced a higher therapist effect, along with suicidality, substance

abuse and depression. Mania produced the lowest therapist effect, which may reflect its relation to general health. A limitation of the study was that not using random slopes in the analysis may have missed those therapists who were better at treating patients of a specific level of severity (e.g., mild or severe).

Nissen-Lie et al. (2016) investigated whether outcome measures and therapist effects were consistent across two different treatment contexts. Data from a US university counseling center and a secondary care unit in Sweden were analysed using the Outcome Questionnaire-45 (OQ-45; Lambert et al. 2004) and the Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM; Evans et al. 2002) respectively. MLM showed that therapists effective in one domain of an outcome measure tended also to be effective in other domains, a finding that held across both treatment centres. However, in the US sample there were no therapist effects found for the OQ-45, whereas in the Swedish sample therapist effects for the CORE-OM ranged from 5.7% to 10%. It is likely that the differences in the severity of patients between the two centers will have contributed to the difference in effects. However, the authors also attributed the assignment of patients to therapist, based on CORE-OM scores, at the Swedish center, as a possible cause. The extent and methods of patient allocation is often unreported or unclear in studies, yet it may have some effect on therapist outcome variability.

Schiefele et al. (2017) combined data from eight naturalistic datasets and used standardized outcomes and MLM, controlling for intake severity, to find an overall therapist effect of 6.7%. Individual therapist effects across the datasets ranged from 2.7–10.2%, with a weighted average of 5.7%. The authors identified the heterogeneity of the studies as a reason for the range of effects. The study also provided sample size recommendations for the number of therapists and number of patients per therapist required for practice-oriented studies.

Specialist/focused settings.

Two studies presented with substantially different characteristics from those reported in the preceding three groups. These two studies (Laska, Smith, Wislocki, Minami, & Wampold 2013; Wiborg, Knoop, Wensing, & Bleijenberg 2012) focused on highly specific patient presenting problems that targeted psychological/physical issues and using outcome measures not adopted in any other studies. Laska et al. (2013) drew on an archival dataset of veterans and utilised supervisor ratings of therapist characteristics, similarly to Green et al. (2014). A therapist effect of 12% was found. Supervisors identified characteristics of more effective therapists including the ability to

address, in particular, client avoidance, adopt a flexible interpersonal style and the ability to build strong therapeutic alliances.

Wiborg et al. (2012) investigated therapist effects in manualised CBT for chronic fatigue syndrome at three community-based mental health care centres. A therapist effect of 21% was found in terms of post-treatment fatigue. This therapist effect decreased when therapists had a more negative attitude towards use of evidence-based treatment manuals. It was also found that the setting in which therapy was delivered had an effect on outcomes, with negative attitudes towards manualization being more clustered within certain treatment centers.

Discussion

This review has provided a systematic examination and evaluation of the status of therapist effects research for the period 2012 to 2016 inclusive, as well as determining, as recommended by Baldwin and Imel, whether study size has increased and whether studies have been specifically designed to address therapist effects. We found studies reported therapist effects in 19 of the 20 studies meeting the inclusion criteria, confirming previous evidence that differences in the effectiveness of therapists occurs across a wide range of clinical settings, patient groups, and also across datasets drawn from routine practice or trials (Baldwin & Imel 2013; Crits-Christoph et al. 1991). Indicative of this variability in setting and design is the range in the size of therapist effects found, namely 0.2–29.0%. However, this range was narrower than the 0–48.7% range reported by Crits-Christoph et al. (1991). The current finding of a weighted average therapist effect of 5% across 31 models lies within the average range of 3–7% reported by Baldwin and Imel (2013).

Although a 5% effect is small relative to the effect of patient variability, studies from the review reported some therapists being consistently more than twice as effective as others after controlling for case-mix (e.g., Firth et al. 2015; Saxon & Barkham 2012). This review confirms that therapists make an important contribution to the variability in patient outcomes (Baldwin & Imel 2013). More specifically, patient intake severity is emerging as a consistent predictor of therapist effect size, with larger effects occurring with more severe patients. This effect was observed in the four clinical groupings of studies (i.e., with university counseling centers at the lower end and specialist/focused settings at the upper end), but this remains an observation in terms of clinical settings in the absence of a greater number of studies within each of the groups. There would, however, be a logic in therapist effects being potentially more critical in clinical settings where the patient population is more severe as a parallel to the finding that the more severe the patient, the more it matters which therapist a patient sees (Schiefele et al. 2017).

However, an overall therapist effect of 5% masks the differences in the size of the effect found in each study. These differences arise from a combination of design and context factors that have the potential to decrease or increase the size of the effect. The analytic method and the sample size are key factors to the size and reliability of therapist effects. The most consistent analytic methods and largest samples were found in naturalistic studies and the therapist effects were also more consistent, indicating the importance of methods and an appropriate sample size in reported effects. However, some of these groupings contain few studies and there is heterogeneity between the studies in the same group. Therefore, this interpretation is tentative.

The finding of larger therapist effects in RCTs compared to naturalistic studies is contrary to the results reported by Baldwin and Imel (2013) and is somewhat counterintuitive. It would be expected that the use of a treatment protocol and the endeavor to ensure that only the treatments differed would yield smaller therapist effects. The current finding may reflect the heterogeneity of RCT studies and limited sample sizes. In the current review, only three trials were quantitatively analysed, and each study had much smaller sample sizes of therapists than recommended by Maas and Hox (2005). In general, the concept of therapist effects in the contexts of an RCT, where variability is suppressed, needs to be carefully considered as compared with practice-based studies where variability is a natural component. However, the fact that therapist effects have been found in RCTs provides further evidence as to the prevalence of the phenomenon. But future trials need to be designed using more therapists to achieve the required number of patients in order to better understand therapist effects in each treatment arm of a trial.

Naturalistic studies, with their larger sample sizes, particularly of therapists, appear better suited to the study of therapist effects because they allow for suitably powered MLM analyses. Indeed, in only the two specialist/focused studies (Laska et al. 2013; Wiborg et al., 2012) and one of the naturalistic studies (Hayes et al. 2015) did the number of patients not meet the criterion of 1200, the product of the number of therapists and the mean number of patients per therapist, as recommended by Schiefele et al. (2017). However, the mean number of patients per therapist only provides an average and does not indicate the lowest number of patients allocated to a therapist, which provides confidence in the value of the patients per therapist calculation. Using the lowest number of patients assigned to a therapist as representing the most conservative number of patients per therapist in any single study and using this as the multiplier with the number of therapists, nine of the 17 naturalistic studies met the criterion of 1200 patients as proposed by Schiefele and colleagues (including the Schiefele et al., study). Hence, fractionally in excess of half the practice-based studies could confidently be said to have met the guideline criterion proposed by Schiefele et al. (2017).

Regarding Baldwin and Imel's (2013) call to increase sample sizes of studies, the number of patients, in particular, increased in the sample of current studies. But recall that in Baldwin and Imel's (2013) review, the median number of therapists per study was only 9 and in only two (4%) of 46 studies did the number of patients per therapist exceed 30. In the current sample comprising RCTs and practice-based studies, the median number of therapists per study was 57.5 and in 12 (60%) of the studies the number of patients per therapist exceeded 30 Hence, it would appear that the recommendation has been heeded and sample sizes have increased, although this applies more to practice-based studies rather than RCTs.

However, an issue remains as to the extent to which studies, even practice-based studies, are designed as therapist effect studies. Studies of routinely collected data remain limited to the variables already collected, and few contain much more than the basic therapist variables. The primary hallmark of studies is the large N of patients. Information that may provide insight into therapist effects is rarely available. However, a small number of studies have applied different methods to move the study of therapist effects forward, by linking routinely collected data to therapist questionnaires of specific variables that might impact on the effect (e.g. Green et al. 2014). A recommendation we would make is that this area of work requires specific studies of therapist effects that collect multiple measures on a sufficient number of therapists (e.g., minimum of 50) as well as meeting the target of a minimum of 1200 patients overall (Schiefele et al. 2017). And, consistent with this approach, that reports should include the same level of information on therapists as for patients. The reporting should also display the actual distribution of individual patients to therapists.

A theme across the studies in the current review was that the more complex the outcome measure (i.e., broader based sampling symptoms, functioning, relationships), the higher the therapist variability – again reflecting findings of the influence of severity on therapist effect (Saxon & Barkham 2012). However, a recommendation from the current study is that reliable and well validated outcome measures are used, and that the reliability of measures is reported, particularly where the outcome used is a subscale (e.g. Owen et al. 2016). Some of the studies in the current review (e.g., Kraus et al. 2016; Nissen-Lie et al. 2016) explicitly used subscales in pursuit of more specific effects. However, many of the subscales of such instruments used (e.g., CORE-OM, TOP) lack evidence of discriminant validity between subscales. In future, we recommend that studies report evidence of the discriminant validity if multiple outcomes are used, otherwise each subscale is not yielding reliable additional information.

Despite the identification of some patterns of therapist effect size, the overriding observation from this review was the degree of heterogeneity of studies, particularly in terms of important factors in determining therapist variability, such as study populations and outcomes and sample sizes at different levels. This concern regarding heterogeneity dissuaded us from carrying out a meta-analysis. Indeed, we suggest that the state of current research argues against carrying out a meaningful meta-analysis. It may be more profitable to determine why there are differences between therapist effects reported across studies rather than to attempt to determine a single point estimate for therapist effects.

This review, by identifying some of the potential causes of heterogeneity aims to inform future study designs. It is also worth noting that a smaller therapist effect is not an indicator of generally better outcomes, only that there is less variability around the average therapist outcomes. For patients, a smaller therapist effect indicates a more restricted range of possible patient outcomes, but it does not state what the absolute value of an outcome might be. The aim is to improve the outcome represented by the 'average' therapist and the study of variability and therapist effects is a means to achieve this aim.

Study limitations

There are a number of caveats to the present review, many of which are due to the limitations of the included studies and their reported descriptions and results. Where it was reported, studies varied in the extent to which patients were randomised or allocated to therapists, and the effect of method of patient allocation on the size of therapist effects is currently unclear (e.g. Goldsmith et al. 2015; Moyers et al. 2016; Nissen-Lie et al. 2016).

There is a tendency to assume the reliability, validity and meaning of both the therapist and patient measures used in therapist effect studies. Different care systems will also dictate differing methods, timings and intensity of data collection and studies tend not to report the timing of measurements and how this is integrated into psychological care. For example, taking an outcome measure following a session is different to taking the same measure before it.

For the current review, the same search terms as Baldwin and Imel (2013) were used, which may not have identified all recent studies. For example, a specific term searching for counseling was not used. Also, stringent inclusion and exclusion criteria limited studies to those that specifically focused on therapist effects, and predominantly on outcome measures. Importantly, we excluded 11 studies that were judged to focus on process variables (e.g., alliance). Applying the 5-point scale post hoc to

determine power within each study (i.e., N of therapists and patients), we found all studies to be rated 2 with the exception of Imel et al. (2014), which was rated 4 due to employing 189 therapists but the patient sample included standardized patients with a mean of less than 3 sessions per therapist. In sum, none of these studies met the criterion of 1200 patients (Schiefele et al., 2017). Inclusion of any of these studies may have impacted on the results. For example, Huppert et al. (2014) found small effects in the context of a trial, which would likely have reduced the therapist effect reported for trial data given the small number of studies reported.

After completing the review and at a stage too late for inclusion in the tables, we found a study by Berglar et al. (2016) that was relevant, within the search time frame, and only listed in Web of Science but which, inexplicably, did not get identified by the search terms and, therefore, did not appear in the original pool of 1566 references. The study comprised 237 patients and 68 therapists, but only one-third saw five or more patients each and no therapist saw 10 patients or more. Hence, it was rated as 3 for power and did not meet the criterion of 1200 patients. Their results supported the phenomenon of a therapist effect and, most interestingly, found that the therapists' impact on treatment outcome not only increased the higher the severity of patients' psychological problems, but that more effective therapists worked even more effectively with patients with higher levels of psychological severity.

The calculation of overall therapist effects, whilst being indicative of general trend, combines data from a range of different contexts and is limited to the particular effects that particular studies reported. For example, some studies accounted for initial severity or case mix in their calculations and others did not. It is also worth noting that whilst large routine datasets can provide sufficiently powered studies, such datasets are predetermined and often driven by pragmatic audit and evaluation concerns rather than theory or research, and it is virtually impossible to include additional measures to extant data collection systems.

Recommendations and implications for practice and policy

This review has shown that the therapist effect reported by Baldwin and Imel (2013) is a durable and robust phenomenon that creates many potential implications for the delivery of services and the training and supervision of therapists. Differential effectiveness is unlikely to be due to the action of a single factor and is far more likely a multicomponent phenomenon. One hypothesis is that it relates to individual therapists carrying out a number of selected elements very well and that these elements differ or overlap between therapists.

However, certain actions for training and practice arise. In terms of training, findings may challenge the ever-increasing focus on academic achievements in the selection of therapists for training and might suggest an emphasis on processes containing active components; for example, role play (Armstrong 2001). Similarly, therapist characteristics such as resilience might also play a role in selection (Green et al. 2014; Pereira et al. 2017).

Once trained, the allocation of patients to therapists should take account of patient severity and more severe patients need to be matched to more effective therapists. Some of the methods developed in therapist effect papers are able to identify the more effective therapists (e.g. Saxon & Barkham 2012). These methods could also enhance clinical supervision by providing more reliable feedback regarding the relative effectiveness of a therapist compared to their peers and include wider indices of patient outcome such as dropout, completion and clinical change rates (e.g. Green et al. 2014; Saxon et al. 2017). Highly effective therapists could be given a clinical supervision role in services and make recordings of sessions regularly available to colleagues and peers. Service managers need to try to encourage an organisational climate that recognises therapist effects without critical judgment (Hunter, Bedell, & Mumford 2007). In this climate, staff can be engaged and curious as to any identified differences and seek methods to close the gap in terms of patient outcomes between therapists via group supervision.

Finally, in terms of policy, psychological treatment guidelines (e.g., the UK National Institute for Health and Care Excellence guidelines) could include statements to the effect that variability exists as to the outcomes achieved by individual therapists. This would emphasise that even within the realm of evidence-based practice, the role of the therapist is still important and would signal to patients that developing and maintaining effective relationships with therapists are paramount.

Recommendations for future therapist effects research

More therapist effect studies and more homogeneous studies are required to produce an overall robust therapist effect. Future research should aim to acquire and analyse the largest samples of practice-based data, in terms of both patients and therapists but particularly therapists, in order to use MLM and produce reliable estimates of effects. One of the largest standardized datasets of routinely collected data from services is that derived from the UK's IAPT initiative and yet while it gathers data on patients and clinical services, it fails to collect or be able to analyse data in relation to therapists and their contribution to patient outcomes. Datasets that are of a greater duration can allow the variability of therapist effects over time to be studied. Studies of therapist effects should use validated and meaningful outcome measures and should also report the manner of patient

allocation and measurement time-points more explicitly. Case-mix and other variables and the levels included in the multilevel analysis should also be reported. But perhaps most crucially, as noted earlier and in order to move towards studies designed as therapist effect studies, more information needs to be collected on a range of therapist variables together with data on personal qualities in addition to more standard information (e.g., number of years' experience, etc.). Very large samples from multiple sites would allow for the study of therapist effects in relation to clinic effects. Also, the role of the clinical supervisor. Hence, there is the potential for four-level models in which patients are nested within therapists, who are nested within clinical supervisors, who are nested within clinics.

Conclusions

Overall, this review has found that across a wide variety of contexts, treatments, outcome measures and patient groups, therapist effects are a significant and robust phenomenon. The average therapist effect found (5%) was within the 3–7% indicated by the previous systematic review (Baldwin & Imel 2013), thereby implying some stability to the therapist effects phenomenon. However, overall there was a large degree of heterogeneity across studies. Although studies are addressing new aspects (e.g., investigating therapist effects over time, low intensity treatments, and comparing outcome measures), reports with sufficient power at both patient and, in particular, therapist levels are clearly still required. The study of therapist effects can be considered a method to better understand therapist variability and, by doing so, to generate research questions with the aim of improving the effectiveness of the 'average' therapist as well as reducing the variability between therapists.

References

Adelson, L.J., & Owen, J., 2012. Bringing the psychotherapist back: Basic concepts of reading articles examining therapist effects using multilevel modelling. *Psychotherapy 49*, 152–162. https://doi.org/10.1037/a0023990.

Ali, S., Littlewood, E., McMillan, D., Delgadillo, J., Miranda, A., Croudace, T., & Gilbody, S., Heterogeneity in patient-reported outcomes following low-intensity mental health interventions: A multilevel analysis, *PLoS One 9* (2014) 1–13. https://doi:10.1371.journal.pone.0099658

Armstrong, J.S., 2001. Role playing: A method to forecast decisions. In: Armstrong, J.S. (Ed.), Principles of forecasting: A handbook for researchers and practitioners. Kluwer. Baldwin, S.A., Imel, Z.E., 2013. Therapist effects: Findings and methods. In: Lambert, M.J. (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change*, 6th ed. Wiley, New York, NY, pp. 258–297.

Barkham, M., Lutz, W., Lambert, M.J., & Saxon, D., 2017. Therapist effects, effective therapists and the law of variability. In: Castonguay, L.G., Hill, C.E. (Eds.), *Therapist effects: Toward understanding how and why some therapists are better than others*. American Psychological Association, Washington, pp. 13–36.

Beutler, L.E., Malik, M., Alimohamed, S., Harwood, T.M., Talebi, H., Noble, S., & Wong, E., 2004. Therapist variables. In: Lambert, M.J. (Ed.), *Bergin & Garfield's handbook of psychotherapy and behavior change*, 5th ed. Wiley, New York, NY, pp. 227–306.

Boswell, J.F., Kraus, D.R., Constantino, M.J., Bugatti, M., & Castonguay, L.G., 2017. The implications of therapist effects for routine practice, policy, and training. In: Castonguay, L.G., Hill, C.E. (Eds.), *Therapist effects: Toward understanding how and why some therapists are better than others.* American Psychological Association, Washington, pp. 309–323.

Brown, G.S., Lambert, M.J., Jones, E.R., & Minami, T., 2005. Identifying highly effective psychotherapists in a managed care environment. *American Journal of Managed Care, 11*, 513–520.

Cella, M., Stahl, D., Reme, S.E., & Chalder, T., 2011. Therapist effects in routine practice: An account from chronic fatigue syndrome. *Psychotherapy Research, 21*, 168–178. https: //doi.org/10.1080/10503307.2010.535571.

Chow, D. L., Miller, S. D., Seidel, J. A., Kane, R. T., Thornton, J. A., & Andrews, W. P. (2015). The role of deliberate practice in the development of highly effective psychotherapists. *Psychotherapy 52,* 337–345. https://doi:10.137/ pst0000015.

Crits-Christoph, P., Baranackie, K., Kurcias, J.S., Beck, A.T., Carroll, K., Perry, K., ... Zitrin, C., 1991. Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research 1*, 81–91. https://doi.org/10.1080/10503309112331335511.

Crits-Christoph, P., Tu, X., & Gallop, R., 2003. Therapists as fixed versus random effects – Some statistical and conceptual issues: A comment on Siemer and Joorman. *Psychological Methods 8*, 518–523. https://doi.org/10.1037/1082-989X.8.4.518.

Dinger, U., Strack, M., Leichsenring, F., Wilmers, F., & Schauenburg, H., 2008. Therapist effects on outcome and alliance in inpatient psychotherapy. *Journal of Clinical Psychology 64*, 344–354. https://doi.org/10.1002/jclp.20443.

Downs, S.H., & Black, N., 1998. The feasibility of creating a checklist for the assessment of the methodological quality of both randomised and non-randomised studies of health care interventions. *Journal of Epidemiological Community Health 52*, 377–384. https: //doi.org/10.1136/jech.52.6.377.

Elkin, I., Falconnier, L., Martinavich, Z., & Mahoney, C., 2006. Therapist effects in the national institute of mental health treatment of depression collaborative research program. *Psychotherapy Research 16*, 144–160. https://doi.org/10.1080/ 10503300500268540.

Erickson, S.J., Tonigan, J.S., & Winhusen, T., 2012. Therapist effects in a NIDA CTN intervention trial with pregnant substance abusing women: Findings from a RCT with MET and TAU conditions. *Alcoholism Treatment Quarterly, 30*, 224–237. https: //doi.org/10.1080/07347324.2012.663295.

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, K., 2002. Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. *The British Journal of Psychiatry 180*, 51–60. https://doi. org/10.1192/bjp.180.1.51.

Firth, N., Barkham, M., Kellett, S., & Saxon, D., 2015. Therapist effects and moderators of effectiveness and efficiency in psychological wellbeing practitioners: A multilevel modelling analysis, *Behaviour Research and Therapy 69*, 54–62, https://doi.org/10. 1016/j.brat.2015.04.001.

Goldberg. S.B., Hoyt, W.T., Nissen-Lie, H.A., Nielsen, S.L. and Wampold, B.E., 2016. Unpacking the therapist effect: Impact of treatment length differs for high- and low-performing therapists, *Psychotherapy Research 12*, 1–13, https:// doi.org/10.1080/10503307.2016.1216625.

Goldberg, S.B., Rousmaniere, T., Miller, S.D., Whipple, J., Nielsen, S.L., Hoyt, W.T. and Wampold, B.E., 2016. Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting, *Journal of Counseling Psychology 63*, 1–11, https://doi.org/10.1037/cou0000131.

Goldsmith, L.P., Dunn, G., Bentall, R.P., Lewis, S.W. and Wearden, A.J., 2015. Therapist effects and the impact of early therapeutic alliance on symptomatic outcome in chronic fatigue syndrome, *PLoS One* 10, 1–13, https://doi.org/10.1371/journal.pone.0144623.

Green, H., Barkham, M., Kellett, S. and Saxon, D., 2014. Therapist effects and IAPT psychological wellbeing practitioners (PWPs): A multilevel modelling and mixed methods analysis, *Behaviour Research and Therapy 63*, 43–54, https://doi.org/10.1016/j. brat.2014.08.009.

Hayes, J.A., McAleavey, A.A., Castonguay, L.G. and Locke, B.D., 2016. Psychotherapists' outcomes with white and racial/ethnic minority clients: First, the good news, *Journal of Counseling Psychology 63*, 261–268, https://doi.org/10.1037/ cou0000098.

Hayes, J.A., Owen, J., & Bieschke, K.J., Therapist differences in symptom change with racial/ethnic minority clients, Psychotherapy 52 (2015) 308–314, https://doi.org/10.1037/a0037957.

Hox, J., 2010. *Multilevel analysis: Techniques and applications*, 2nd ed. Routledge, UK.
18

Hunter, S.T., Bedell, K.E., Mumford, M.D., 2007. Climate for creativity: A quantitative review. *Creativity Research Journal 19*, 69–90. https://doi.org/10.1080/ 10400410709336883.

Huppert, J.D., Bufka, L.F., Barlow, D.H., Gorman, J.M., Shear, M.K., Woods, S.W., 2001. Therapists, therapist variables, and cognitive-behavioural therapy outcome in a multicentre trial for panic disorder. *Journal of Consulting and Clinical Psychology 69*, 747–755. https://doi.org/10.1037//0022-006X.69.5.747.

Huppert, J.D., Kivity, Y., Barlow, D.H., Gorman, J.M., Shear, M.K., Woods, S.W., 2014. Therapist effects and the outcome-alliance correlation in cognitive behavioural therapy for panic disorder with agoraphobia. *Behaviour Research and Therapy 52*, 26–34. https://doi.org/10.1016/j/brat.2013.11.001.

Kazdin, A.E., Bass, D., 1989. Power to detect differences between alternative treatments in comparative psychotherapy outcome research. *Journal of Consulting and Clinical Psychology 57*, 138–147.

Kim, D.-M., Wampold, B.E., Bolt, D.M., 2006. Therapist effects in psychotherapy: A random-effects modelling of the national institute of mental health treatment of depression collaborative research program data. *Psychotherapy Research 16*, 161–172. https: //doi.org/10.1080/10503300500264911.

Kraus, D.R., Bentley, J.H., Boswell, J.F., Constantino, M.J., Baxter, E.E. and Castonguay, L.G., 2016.Predicting therapist effectiveness from their own practice-based evidence, *Journal of Consulting and Clinical Psychology 84*, 473–483, https://doi.org/10.1037/ccp0000083.

Kraus, D.R., Seligman, D.A., Jordan, J.R., 2005. Validation of a behavioral health treatment outcome and assessment tool designed for naturalistic settings: The Treatment Outcome Package. *Journal of Clinical Psychology 61*, 285–314. https://doi.org/10. 1002/jclp.20084.

Lambert, M.J., Morton, J.J., Hatfield, D., Harmon, C., Hamilton, S., Reid, R.C., ... Burlingame, G.B., 2004. *Administration and scoring manual for the Outcome Questionnaire-45*. American Professional Credentialing Services, UT.

Laska, K.M., Smith, T.L., Wislocki, A.P., Minami, T. and Wampold, B.E., 2013. Uniformity of evidence-based treatments in practice? Therapist effects in the delivery of cognitive processing therapy for PTSD, Journal of Counseling Psychology 60, 31–41, https://doi.org/10.1037/aa0031294.

Lutz, W., Barkham, M., 2015. Therapist effects. *The encyclopedia of clinical psychology*. Blackwell: Wiley.

Maas, C.J.M., Hox, J.J., 2005. Sufficient sample sizes of multilevel modelling. *Methodology 1*, 86–92. https://doi.org/10.1027/1614-1881.1.3.86.

Martindale, C., 1978. The therapist as fixed-effect fallacy in psychotherapy research. *Journal of Consulting and Clinical Psychology 46*, 1526–1530.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine 151*, 264–269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135.

Moyers, T.B., Houck, J., Rice, S.L., Longabaugh, R. and Miller, W.R., 2016. Therapist empathy, combined behavioural intervention, and alcohol outcomes in the COMBINE research project, *Journal of Consulting and Clinical Psychology 84*, 221–229, https://doi.org/10.1037/ccp0000074.

Najavits, L.M., Strupp, H.H., 1994. Differences in the effectiveness of psychodynamic therapists: A process-outcome study. *Psychotherapy 31*, 114–123.

Nissen-Lie, H.A., Goldberg, S.B., Hoyt, W.T., Falkenström, F., Holmqvist, R., Nielsen, S.L. and Wampold, B.E., 2016. Are therapists uniformly effective across patient outcome domains? A study on therapist effectiveness in two different treatment contexts, *Journal of Counseling Psychology 63,* 367–378, https://doi. org/10.1037/cou0000151.

Okiishi, J.C., Lambert, M.J., Eggett, D., Nielsen, L., Dayton, D.D., Vermeersch, D.A., 2006. An analysis of therapist treatment effects: Toward providing feedback to individual therapists on their clients' psychotherapy outcome*. Journal of Clinical Psychology 62*, 1157–1172. https://doi.org/10.1002/jclp.

Okiishi, J.C., Lambert, M.J., Nielsen, L., Ogles, B.M., 2003. Waiting for supershrink: An empirical analysis of therapist effects. *Clinical Psychology & Psychotherapy 10*, 361–373. https://doi.org/10.1002/cpp.383.

Owen, J., Drinane, J.M., Idigo, K.C., Valentine, J.C., 2015. Psychotherapist effects in meta-analyses: How accurate are treatment effects? *Psychotherapy 52*, 321–328. https://doi.org/10.1037/pst0000014.

Owen, J.J., Adelson, J., Budge, S., Kopta, S.M. and Reese, R.J., 2016. Good-enough level and dose-effect models: Variation among outcomes and therapists, *Psychotherapy Research 26*, 22–30, https://doi.org/10.1080/10503307.2014. 966346.

Pereira, J-A., Barkham, M., Kellett, S. and Saxon, D.,2017. The role of practitioner resilience and mindfulness in effective practice: A practice-based feasibility study, *Administration and Policy in Mental Health and Mental Health Research 44*, 691–704, https://doi.org/10.1007/s10488-016-0747-0.

Raudenbush, S.W., Bryk, A.S., 2002. *Hierarchical linear models: Applications and data analysis methods*. Springer, Thousand Oaks, CA.

Ricks, D.F., 1974. Supershrink: Methods of a therapist judged successful on the basis of adult outcomes of adolescent patients. In: Ricks, D.F., Roff, M., Thomas, A. (Eds.), *Life history research in psychopathology*. Vol. 3, University of Minnesota Press, Minneapolis, pp. 275–297.

Saxon, D. and Barkham, M., 2012. Patterns of therapist variability: Therapist effects and the contribution of patient severity and risk, *Journal of Consulting and Clinical Psychology 80*, 535–546, https://doi.org/10.1037/a0028898.

Saxon, D., Firth, N. and Barkham, M., 2017. The relationship between therapist effects and therapy delivery factors: Therapy modality, dosage, and non-completion, *Administration and Policy in Mental Health and Mental Health Research 44*, 705–715, https://doi.org/10.1007/s10488-016-0750-5.

Schiefele, A-K., Lutz, W., Barkham, M. Rubel, J., Böhnke, J., Delgadillo, J., ... Lambert, M.J., 2017. Reliability of therapist effects in practice-based psychotherapy research: A guide for the planning of future studies, *Administration and Policy in Mental Health and Mental Health Research 44*, 598–613, https://doi.org/10.1007/ s10488-016-0736-3.

Wampold, B., 2001. In: Mahwah, N.J. (Ed.), *The great psychotherapy debate: Models, methods, and findings*. Lawrence Erlbaum Associates Inc.

Wampold, B.E., 2005. What should be validated? The psychotherapist. In: Norcross, J.C., Beutler, L.E., Levant, R.F. (Eds.), *Evidence-based practices in mental health: Debate and dialogue on the*

fundamental questions. American Psychological Association, Washington DC, pp.200-208, 236-238.

Wampold, B.E., 2007. Psychotherapy: The humanistic (and effective) treatment. *American Psychologist 62*, 857–873. https://doi.org/10.1037/0003-066X.62.8.857.

Wampold, B.E., Baldwin, S.A., Grosse Holtforth, M., Imel, Z.E., 2017. What characterizes effective therapists? In: Castonguay, L.G., Hill, C.E. (Eds.), Therapist effects: *Toward understanding how and why some therapists are better than others*. American Psychological Association, Washington, pp. 37–53.

Wampold, B.E., Brown, G.S., 2005. Estimating variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. *Journal of Consulting and Clinical Psychology 73*, 914–923. https://doi.org/10.1037/0022-006X.73.5.914.

Wampold, B.E., Imel, Z.E., 2015. *The great psychotherapy debate: The evidence for what makes psychotherapy work*, 2nd ed. Routledge, New York, NY.

Wiborg, J.F., Knoop, H., Wensing, M. and Bleijenberg, G., 2012. Therapist effects and the dissemination of cognitive behavior therapy for chronic fatigue syndrome in community-based mental health care, *Behaviour Research and Therapy 50*, 393–396, https://doi.org/10.1016/j.brat.2012.03.002.

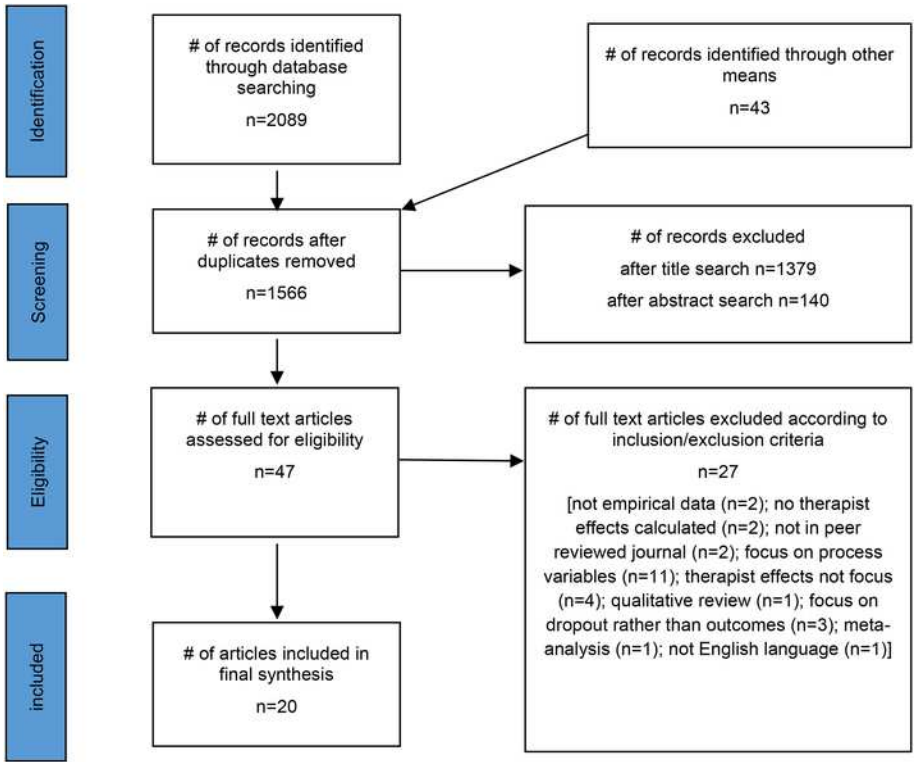

Figure 1: PRISMA diagram of study selection process

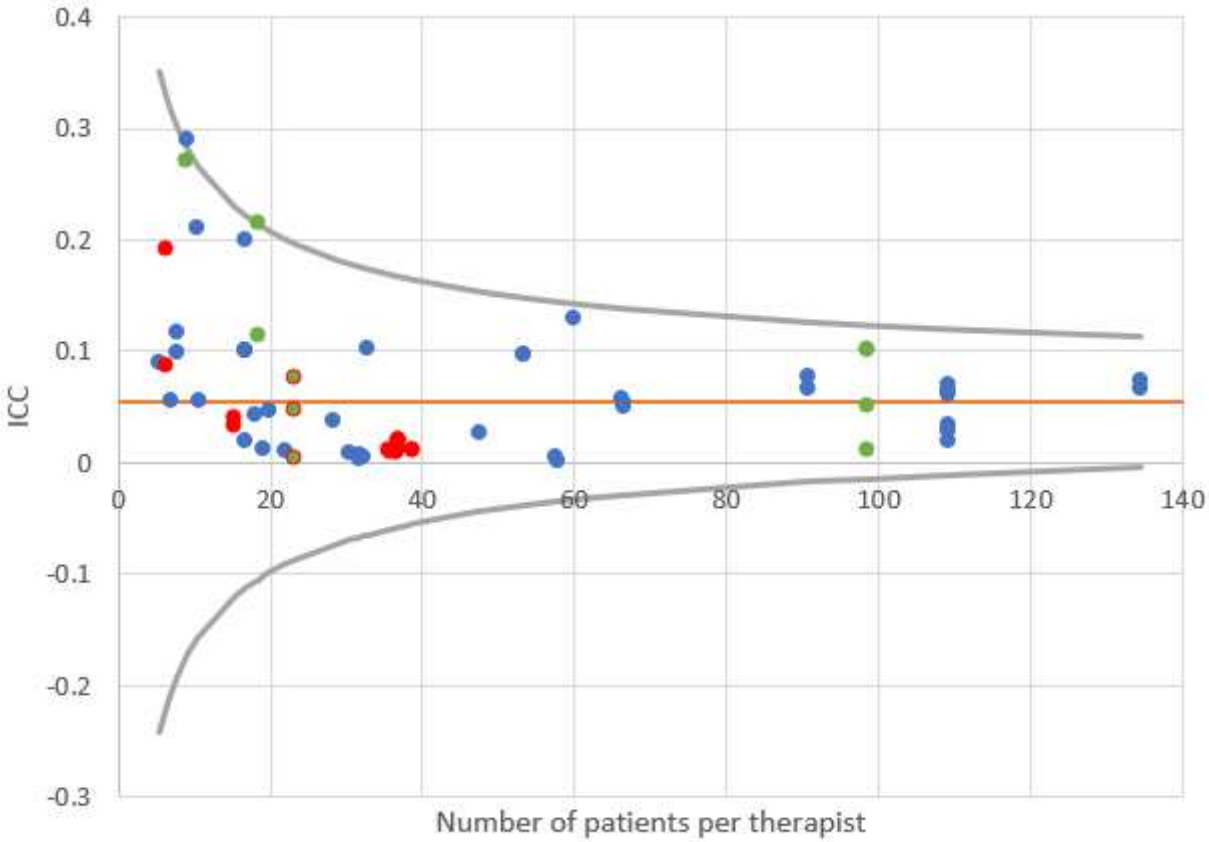

Figure 2. Funnel plot of ICCs for all models. *Note:* Orange line indicates overall weighted mean by number of patients per therapist; each dot represents a model from a review study; green denotes RCT, red denotes UCC, blue denotes other naturalistic (non-UCC) studies (see Tables 2 and 3).

Table 1

Summary of therapist effects study characteristics

	No. of patients	No. of therapists	Mean patients per therapist	SD	Lowest patients per therapist	Highest patients per therapist	Diagnosis	Outcome measure(s)	Intervention	Treatment center(s)	Therapist effects analysis[2]	Significant therapist effects found	Quality checklist rating
RCT studies													
Erickson et al. (2012)	91	10	9	n/g	5	n/g	Substance abuse	ASI-Lite; URICA; HAq-II	Motivational Enhancement therapy	Community outpatient centers	GLM/linear regression/ HLM[3]	Yes	20
Goldsmith et al. (2015)	196 [+ 100 control patients not analyzed]	3	65.3	n/g	64‡	66	Chronic fatigue syndrome	Chalder fatigue scale; SF-36	Pragmatic rehabilitation; supportive counseling	Primary care center	Regression	No	23
Moyers et al. (2016)	700	38	18.4	14.1	1	47	Alcohol-related difficulties	PDA; DDD	Behavioral therapy	Alcohol treatment centers	MLM	Yes	24
Practice-based studies													
University counseling centers													
Goldberg, Hoyt et al. (2016)	5828*	158††	36.9	47.8	10‡	333	Mixed	OQ-45	Mixed psychotherapy	University counseling center	MLM	Yes	27
Goldberg, Rousmaniere et al. (2016)	6591*	170††	38.8	51.4	10‡	360	Mixed	OQ-45	Mixed psychotherapy	University counseling center	MLM	Yes	27
Hayes et al. (2015)	228	36	6	n/g	4	13	Depression/ anxiety/relation-ship issues/ academic distress	OQ-45	Mixed psychotherapy	University counseling center	MLM	Yes	20

Table 1 continued

	No. of patients	No. of therapists	Mean patients per therapist	SD	Lowest patients per therapist	Highest patients per therapist	Diagnosis	Outcome measure(s)	Intervention	Treatment center(s)	Therapist effects Analysis[3]	Significant therapist effects found	Quality checklist rating
Hayes et al. (2016)	3825*	251††	15.3	10	6	72	Mixed	CCAPS-62	Mixed psychotherapy/ counseling	University counseling centers	MLM	Yes	24
Owen et al. (2016)	13664	586††	23.3	n/g	2	455	Mixed	BHM-20	Mixed psychotherapy	University counseling centers	MLM	Yes	22
Primary care													
Ali et al. (2014)	1376	38	36.2	25.5	1	109	Depression/ anxiety	PHQ-9; GAD-7	Brief low-intensity therapy	Primary care IAPT service[2]	HLM[4]	Yes	26
Firth et al. (2015)	6111*	56†	109	n/g	30‡	n/g	Depression/ anxiety	PHQ-9; GAD-7; WSAS	Low intensity therapy	Primary care IAPT service[2]	MLM	Yes	26
Green et al. (2014)	1122	21	53.6	n/g	8	197	Depression/ anxiety	PHQ-9; GAD-7	Guided self-help	Primary care IAPT services[2]	MLM	Yes	23
Pereira et al. (2017)[1]	4980	37	134.6	100.1	24‡	n/g	Depression	PHQ-9; WSAS; IMD	CBT/counseling & low-intensity therapy	Primary care IAPT service[2]	MLM	Yes	24
Saxon & Barkham (2012)	10786*	119††	91	n/g	30‡	n/g	Depression/ anxiety	CORE-OM	CBT, counseling	Primary care psychotherapy services[2]	MLM	Yes	27
Saxon et al. (2017)[1]	4034*	61†	66.1	n/g	20‡	n/g	Depression/ anxiety	PHQ-9	Mixed	Primary care IAPT service[2]	MLM	Yes	26

Table 1 continued

	No. of patients	No. of therapists	Mean patients per therapist	SD	Lowest patients per therapist	Highest patients per therapist	Diagnosis	Outcome measure(s)	Intervention	Treatment center(s)	Therapist effects Analysis[3]	Significant therapist effects found	Quality checklist rating
Mixed													
Chow et al. (2015)	4580	69†	66.4	70.0	10‡	335	Depression/ anxiety	CORE-OM	Mixed psychotherapy	Voluntary (42%); independent practice (39.1%); primary care (8.7%); secondary care (4.3%)	MLM	Yes	24
Kraus et al. (2016)	3540*	59†	60	n/g	60‡	n/g	Mixed	TOP	Psychotherapy	Mixed (outpatient therapy services; independent practice; hospitals; residential settings; day treatment programs)	HLM[4]	Yes	23
Nissen-Lie et al. (2016)	6444*	196††	36.9	47.8	10‡	333	Mixed	OQ-45; CORE-OM	Mixed psychotherapy	University counseling center; primary and secondary care unit	MLM	Yes	25
Schiefele et al. (2017)[1]	48648*	1800††	27	n/g	2	400	Mixed	BSI; BHM-20; MHI; OQ-45; CORE-OM; PHQ-9	Mixed	Mixed	MLM	Yes	26

Table 1 continued

	No. of patients	No. of therapists	Mean patients per therapist	SD	Lowest patients per therapist	Highest patients per therapist	Diagnosis	Outcome measure(s)	Intervention	Treatment center(s)	Therapist effects Analysis[3]	Significant therapist effects found	Quality checklist rating
Specialist/ focused settings													
Laska et al. (2013)	192	25	8	n/g	1	62	PTSD	PCL	Cognitive processing therapy	Veterans hospital – outpatient and community	MLM	Yes	22
Wiborg et al. (2012)	103	10	10	n/g	4	17	Chronic fatigue syndrome	CIS (fatigue subscale)	Manualised CBT for chronic fatigue syndrome	Community-based mental healthcare centers	Random effects modeling[4]	Yes	22

Note. ASI-Lite = Addiction Severity Index-Lite; BAI = Beck Anxiety Inventory; BDI = Beck Depression Inventory; BHM-20 = Behavioral Health Measure-20; BSI = Brief System Inventory; CCAPS-62 = Counseling Center Assessment of Psychological Symptoms-62; CIS = Checklist Individual Strength; CORE-OM = Clinical Outcomes in Routine Evaluation-Outcome Measure; DDD = drinks per drinking day; GAD-7 = Generalised Anxiety Disorder-7; HAq-II = Revised Helping Alliance Questionnaire; IAPT = Improving Access to Psychological Therapies; IMD = Index of Multiple Depravation; MHI = Mental Health Index; OQ-45 = Outcome Questionnaire-45; PCL = PTSD Checklist; PDA = per cent days abstinent; PHQ-9 = Patient Health Questionnaire-9; PTSD = Post-traumatic stress disorder; PDS = Posttraumatic Diagnostic Scale; RCT = Randomised Control Trial; SF-36 = Short Form Health Survey; TOP = Treatment Outcome Package; URICA = University of Rhode Island Change Assessment; WSAS = Work and Social Adjustment Scale; [1]published online in 2016 and thus included in review period; [2] The term 'service' is used in the UK and approximates the US term clinic or center; [3]analysis as reported in the study; [4]alternative term for MLM (Adelson & Owen, 2012). * = N of patients > 1200 as a product of N of therapists x lowest N of patients per therapist; †† = N of therapists > 100; † = N of therapists > 50; ‡ = N of patients per therapist ≥ 10 for all therapists.

Table 2

Summary of methods of study design

	When measures taken (e.g. pre-post/sessional)	Follow-up measures	Measures are symptom-based or broader	Controlled for any aspects (e.g. severity)	How dealt with missing data	Therapists described in enough detail	Primary or secondary study	Other aspects of outcome (e.g. dropout)	How ICC calculated
RCT studies									
Erickson et al. (2012)	Post	1-month, 3-month	Symptom-based	Severity	Excluded from analysis	Substance abuse counselors	Primary	Alliance	GLM repeated measures
Goldsmith et al. (2015)	Pre, post	1-year	Symptom-based (health)	Severity	Weighting	Nursing practitioners	Primary	Alliance	Regression (ANCOVA)
Moyers et al. (2016)	Pre, post	Weeks 8, 16, 26, 52, 68	Symptom-based (drinking outcomes)	none	n/a	Psychologists, counselors, social workers, other behavioural health professionals	Secondary	Empathy	2-levels: patients within therapists
Practice-based studies									
University counseling centers									
Goldberg, Hoyt et al. (2016)	Sessional	n/a	Broad measure of outcomes	Case mix	Not specified	Licensed and trainee therapists	Primary	None	2-levels: patients within therapists
Goldberg, Rousmaniere et al. (2016)	Pre, post	n/a	Broad measure of outcomes	Time and cases used as predictors	Not specified	Licensed and trainee therapists	Primary	Therapist experience	2-levels: patients within therapists
Hayes et al. (2015)	Sessional	n/a	Broad measure of outcomes	Race fixed and varied	Not specified	Counseling and counseling psychology trainees	Primary	None	2-levels: patients within therapists
Hayes et al. (2016)	Pre, post	None	Broad measure of outcomes	Severity	Not specified	Psychologists and psychology trainees	Primary	None	2-levels: patients within therapists
Owen et al. (2016)	Sessional	None	Subscale of broad measure of outcomes	Severity	Excluded from analysis	Psychologists, counselors, psychiatrists, social workers	Primary	None	3-levels: observations within patients within therapists

Table 2 (continued)

	When measures taken (e.g. pre-post/sessional)	Follow-up measures	Measures are symptom-based or wider	Controlled for any aspects (e.g. severity)	How dealt with missing data	Therapists described in enough detail	Primary or secondary study	Other aspects of outcome (e.g. dropout)	How ICC calculated
Primary care									
Ali et al. (2014)	Sessional		Symptom-based	Age, gender, visit number and duration	Not specified	Low intensity therapists	Primary	None	3-levels: sessions within patients within therapists; sensitivity analysis with 2-levels: patients within therapists
Firth et al. (2015)	Sessional	n/a	Symptom-based and functional impairment/deprivation	Severity, demographics (e.g. age, deprivation, employment status, gender), intervention non-completion, no. of sessions	Excluded from analysis	Psychological wellbeing practitioners	Primary	None	2-levels: patients within therapists
Green et al. (2014)	Pre, post	n/a	Symptom-based		Excluded from analysis	Psychological wellbeing practitioners	Primary	Therapist ego strength, intuition, resilience	2-levels: patients within therapists
Pereira et al. (2017)	Pre, post	None	Symptom-based	Severity and case mix	Excluded from analysis	Psychological wellbeing practitioners, CBT therapists, counselors	Primary	Therapist resilience and mindfulness	2-levels: patients within therapists
Saxon & Barkham (2012)	Pre, post	n/a	Broad measure of outcomes	Severity and case-mix	Excluded from analysis	Counselors and psychological therapists	Primary	None	2-levels: patients within therapists
Saxon et al. (2017)	Pre, post	n/a	Symptom-based	Severity and case mix	Excluded from analysis	CBT therapists and counselors	Primary	None	2-levels: patients within therapists

Table 2 (continued)

	When measures taken (e.g. pre-post/sessional)	Follow-up measures	Measures are symptom-based or wider	Controlled for any aspects (e.g. severity)	How dealt with missing data	Therapists described in enough detail	Primary or secondary study	Other aspects of outcome (e.g. dropout)	How ICC calculated
Mixed settings									
Chow et al. (2015)	Pre, post	None	Symptom-based and measure of psychotherapist involvement in deliberate practice	Severity	Not specified	Psychotherapists, psychologists, social workers, marriage and family therapists, counsellors	Primary	Therapist involvement in deliberate practice	3-levels: patients within therapists within organisations
Kraus et al. (2016)	Pre, post	30-180 days	Broad measure of outcomes	Risk	Excluded from analysis	Social workers, counsellors, psychologists, drug/alcohol counsellors, psychiatrists	Primary	None	2-levels: patients within therapists
Nissen-Lie et al. (2016)	Sessional	None	Subscales of broad measures of outcomes	Severity	Not specified	Psychotherapists	Primary	None	2-levels: observations within patients to obtain change measure, then 2-levels: patients within therapists, confirmatory factor analytic model
Schiefele et al. (2017)	Pre, post	n/a	Mixed	Mixed	Mixed	Mixed	Secondary	n/a	2-levels: patients within therapists
Specialist/focused settings									
Laska et al. (2013)	Pre, post	None	Symptom-based	Severity	Excluded from analysis	Psychologists, social workers & trainees	Primary	None	2-levels: patients within therapists
Wiborg et al. (2012)	Pre, post	None	Symptom-based	Severity	Excluded from analysis	CBT therapists	Primary	Therapist attitude towards manualization	

Table 3

Reported ICC values for RCT studies

Author(s) and Date	Conditions for model	ICC	95% CI	No. of patients	No. of therapists	Mean ICC: different measures or subscales	Mean overall ICC based on ICC values in column 3
Erickson et al. (2012)	Substance use – all	.270	n/g	91	10	-	.280
	Substance use – MET condition	.290	n/g	91	10	-	
Goldsmith et al. (2015)	Chalder fatigue – PR	.100	n/g	296	3	.100	.065
	Chalder fatigue – SL	.100	n/g	296	3		
	SF-36 – PR	.050	n/g	296	3	.025	
	SF-36 – SL	.010	n/g	296	3		
Moyers et al. (2016)	Drinking outcomes – untransformed	.214	.108-.338	700	38	-	.164
	Drinking outcomes – log transformed	.114	.029-.221	700	38	-	
Mean ICC, weighted for no. of patients							.129
Mean ICC, weighted for no. of therapists							.174
Mean ICC, weighted for no. of patients per therapist							.082

Note. CI = confidence interval; ICC = Intraclass correlation co-efficient**;** MET = motivational enhancement therapy; n/g = not given; PR = Pragmatic Rehabilitation; RCT = Randomized Control Trial; SF-36 = Short Form Health Survey; SL = supportive listening

Table 4
Reported ICC values for practice-based studies

Author(s) and Date	Conditions for model	ICC	95% CI	No. of patients	No. of therapists	Mean ICC: different measures or subscales	Mean overall ICC based on ICC values in column 3
University counseling							
Goldberg, Hoyt et al. (2016)	OQ-45 - no predictors*	.009	n/g	5794	158	-	.009
	OQ-45 - controlled for case mix (average)*	.009	n/g	5794	158	-	
Goldberg, Rousmaniere et al. (2016)	OQ-45 – time as predictor*	.010	n/g	6591	170	-	.011
	OQ-45 – cases as predictor*	.011	n/g	6591	170	-	
Hayes et al. (2015)	OQ-45 – race fixed*	.087	n/g	228	36	-	.139
	OQ-45 – race varied*	.191	n/g	228	36	-	
Hayes et al. (2016)	CCAPS-62 (DI)*	.039	n/g	3825	251	-	.036
	CCAPS-62 (DI) – controlled for pre-treatment score*	.032	n/g	3825	251	-	
Owen et al. (2016)	BHM-20 – wellbeing*	.004	n/g	13664	586	.004	.042
	BHM-20 - symptom distress*	.046	n/g	13664	586	.046	
	BHM-20 - life functioning*	.075	n/g	13664	586	.075	
Primary care							
Ali et al. (2014)	PHQ-9	.010	.003-.0038	1359	38	.010	.007
	GAD-7	.009	.002-.0039	1366	38	.009	
	PHQ-9 controlled for age & gender	.004	.000-.0043	1174	37	-	
	GAD-7 controlled for age & gender	.006	.001-.0035	1190	37	-	
	PHQ-9 controlled for visit number & duration	.007	.001-.0048	1174	37	-	
	GAD-7 controlled for visit number & duration	.008	.001-.0043	1127	37	-	
	PHQ-9 full sample	.005	.001-.0024	2190	38	.005	
	GAD-7 full sample	.002	.000-.0054	2197	38	.002	
	PHQ-9 above baseline	.012	.002-.0060	703	37	-	
	GAD-7 above baseline	.011	.002-.0057	811	37	-	

Table 4 (continued)

Author(s) and Date	Conditions for model	ICC	95% CI	No. of patients	No. of therapists	Mean ICC: different measures or subscales	Mean overall ICC based on ICC values in column 3
Firth et al. (2015)	PHQ-9	.028	n/g	6111	56	.028	.046
	GAD-7	.019	n/g	6111	56	.019	
	WSAS	.034	n/g	6111	56	.034	
	PHQ-9 – controlled for case mix	.064	n/g	6111	56	-	
	GAD-7 – controlled for case mix	.061	n/g	6111	56	-	
	WSAS – controlled for case mix	.070	n/g	6111	56	-	
Green et al. (2014)	PHQ-9	.097	.058-.174	1122	21	.097	.098
	GAD-7	.098	.058-.176	1122	21	.098	
Pereira et al. (2017)	PHQ-9 – controlled for pre-treatment score	.073	n/g	4980	37	-	.070
	PHQ-9 – controlled for pre- treatment score & case mix	.067	n/g	4980	37		
Saxon & Barkham (2012)	CORE-OM – without risk	.078	n/g	10786	119	-	.072
	CORE-OM – controlled for risk	.066	n/g	10786	119	-	
Saxon et al. (2017)	PHQ-9	.058	n/g	4034	61	-	.058
Mixed settings							
Chow et al. (2015)	CORE-10 full sample	.054	n/g	4580	69	-	.052
	CORE-10 controlled for severity	.051	n/g	4580	69	-	
Kraus et al. (2016)	TOP – risk adjusted	.129	n/g	3540	59	-	.129
Nissen-Lie et al. (2016)	OQ-45 – total*	.019	n/g	5828	158	.019	.060
	OQ-45 – symptom distress*	.020	n/g	5828	158	.020	
	OQ-45 – interpersonal relationships*	.013	n/g	5828	158	.013	[.018 OQ-45]
	OQ-45 – social relationships*	.019	n/g	5828	158	.019	
	CORE-OM – wellbeing	.100	n/g	520	31	.100	
	CORE-OM – problems	.100	n/g	520	31	.100	
	CORE-OM – close relationships	.200	n/g	520	31	.200	[.108 CORE-OM]
	CORE-OM – general	.020	n/g	520	31	.020	
	CORE-OM – social	.100	n/g	520	31	.100	

Table 4 (continued)

Author(s) and Date	Conditions for model	ICC	95% CI	No. of patients	No. of therapists	Mean ICC: different measures or subscales	Mean overall ICC based on ICC values in column 3
Schiefele et al. (2017)	BSI	.055	n/g	668	97		.057
	BSI	.090	n/g	636	120	[.067 BSI]	
	BSI	.055	n/g	752	71		
	BHM-20	.038	n/g	11356	401	.038	
	MHI	.047	n/g	1194	60	.047	
	OQ-45	.043	n/g	2561	143	.043	
	CORE-OM	.102	n/g	25842	789	.102	
	PHQ-9	.027	n/g	5639	119	.027	
Specialist/focused settings							
Laska et al. (2013)	PCL – controlled for pre-treatment score	.117	n/g	192	25	-	.108
	PCL – controlled for pre-treatment score – with rating score	.099	n/g	192	25	-	
Wiborg et al. (2012)	CIS – fatigue severity	.210	n/g	103	10	-	.210
Mean ICC, weighted for no. of patients							.047
Mean ICC, weighted for no. of therapists							.048
Mean ICC, weighted for no. of patients per therapist							.050

Note. BHM-20 = Behavioral Health Measure -20; BSI = Brief Symptom Inventory; CCAPS-62 = Counselling Centre Assessment of Psychological Symptoms; CI = confidence interval; CIS = Checklist Individual Strength; CORE-OM=Clinical Outcomes in Routine Evaluation-Outcome Measure; CORE-10= Clinical Outcomes in Routine Evaluation-10; DI = Distress Index; GAD-7 = Generalised Anxiety Disorder-7; IAPT = Improving Access to Psychological Therapies; ICC = Intraclass correlation co-efficient; n/g = not given; MHI = Mental Health Index; OQ-45 = Outcome Questionnaire-45; PCL = PTSD Checklist; PHQ-9 = Patient Health Questionnaire-9; RCT = Randomized Controlled Trial; *university counseling centers

Appendix A – Modified Downs and Black (1998) Quality Checklist – with explanations of modifications

Reporting

1. *Is the hypothesis/aim/objective of the study clearly described?*

Yes	1
No	0

2. *Are the main outcomes to be measured clearly described in the Introduction or Methods section?*

   If the main outcomes are first mentioned in the Results section, the question should be answered no.

Yes	1
No	0

3. *Are the characteristics of the patients included in the study clearly described?*

   In cohort studies and trials, inclusion and/or exclusion criteria should be given. In case-control studies, a case-definition and the source for controls should be given.

Yes	1
No	0

4. *Are the interventions of interest clearly described?*

   Treatments and placebo (where relevant) that are to be compared should be clearly described.

Yes	1
No	0

5. *Are the distributions of principal confounders in each group of subjects to be compared clearly described?*

   A list of principal confounders is provided.

Yes	2
Partially	1
No	0

6. *Are the main findings of the study clearly described?*

   Simple outcome data should be reported for all major therapist effects so that the reader can check the major analyses and conclusions.
   (This question does not cover statistical tests which are considered below).

Yes	1
No	0

7. *Does the study provide estimates of the random variability in the data for the main outcomes?*

In non-normally distributed data the inter-quartile range of results should be reported. In normally distributed data the standard error, standard deviation or confidence intervals should be reported around the therapist effect. If the distribution of the data is not described, it must be assumed that the estimates used were appropriate and the question should be answered yes.

Yes	1
No	0

8. *Have all important adverse events that may be a consequence of the intervention been reported?*

   This should be answered yes if the study demonstrates that there was a comprehensive attempt to measure adverse events. (A list of possible adverse events is provided).

Yes	1
No	0

9. *Have the characteristics of patients lost to follow-up been described?*

   This should be answered yes where there were no losses to follow-up or where losses to follow-up were so small that findings would be unaffected by their inclusion. This should be answered no where a study does not report the number of patients lost to follow-up.

Yes	1
No	0

10. *Have actual probability values been reported (e.g. 0.035 rather than <0.05) for the main outcomes except where the probability value is less than 0.001?*

Yes	1
No	0

External validity

All the following criteria attempt to address the representativeness of the findings of the study and whether they may be generalised to the population from which the study subjects were derived.

11. *Were the subjects asked to participate in the study representative of the entire population from which they were recruited?*

    The study must identify the source population for patients and describe how the patients were selected. Patients would be representative if they comprised the entire source population, an unselected sample of consecutive patients, or a random sample. Random sampling is only feasible where a list of all members of the relevant population exists. Where a study does not report the proportion of the source population from which the patients are derived, the question should be answered as unable to determine.

Yes	1
No	0
Unable to determine	0

12. *Were those subjects who were prepared to participate representative of the entire population from which they were recruited?*

    The proportion of those asked who agreed should be stated. Validation that the sample was representative would include demonstrating that the distribution of the main confounding factors was the same in the study sample and the source population.

Yes	1
No	0
Unable to determine	0

13. *Were the staff, places, and facilities where the patients were treated, representative of the treatment the majority of patients receive?*

    For the question to be answered yes the study should demonstrate that the intervention was representative of that in use in the source population. The question should be answered no if, for example, the intervention was undertaken in a specialist centre unrepresentative of the hospitals most of the source population would attend.

Yes	1
No	0
Unable to determine	0

Internal validity – bias

14. *Was an attempt made to blind study subjects to the intervention they have received?*

    For studies where the patients would have no way of knowing which intervention they received, this should be answered yes.

Yes	1
No	0
Unable to determine	0

15. *Was an attempt made to blind those measuring the main outcomes of the intervention?*

Yes	1
No	0
Unable to determine	0

16. *If any of the results of the study were based on "data dredging", was this made clear?*

    Any analyses that had not been planned at the outset of the study should be clearly indicated. If no retrospective unplanned subgroup analyses were reported, then answer yes.

Yes	1
No	0
Unable to determine	0

17. *In trials and cohort studies, do the analyses adjust for different lengths of follow-up of patients, or in case-control studies, is the time period between the intervention and outcome the same for cases and controls?*

    Where follow-up was the same for all study patients the answer should yes. If different lengths of follow-up were adjusted for by, for example, survival analysis the answer should be yes. Studies where differences in follow-up are ignored should be answered no.

Yes	1
No	0
Unable to determine	0

18. *Were the statistical tests used to assess the therapist effects appropriate?*

    Were the data analysed within a hierarchical structure (e.g. using Multilevel Modelling), using random effects analysis, or at least involved calculation of the intraclass coefficient (ICC) for therapists?

Yes	1
No	0
Unable to determine	0

19. *Was compliance with the intervention/s assessed?*

    Where there was non compliance with the allocated treatment or where there was contamination of one group, the question should be answered no. For studies where the effect of any misclassification was likely to bias any association to the null, the question should be answered yes.

Yes	1
No	0
Unable to determine	0

20. *Were the main outcome measures used accurate (valid and reliable)?*

    For studies where the outcome measures are clearly described, the question should be answered yes. For studies which refer to other work or that demonstrates the outcome measures are accurate, the question should be answered as yes.

Yes	1
No	0
Unable to determine	0

Internal validity - confounding (selection bias)

21. *Were the patients in different intervention groups (trials and cohort studies) or were the cases and controls (case-control studies) recruited from the same population?*

    For example, patients for all comparison groups should be selected from the same hospital. The question should be answered unable to determine for cohort and casecontrol studies where there is no information concerning the source of patients included in the study.

Yes	1
No	0
Unable to determine	0

22. *Were study subjects in different intervention groups (trials and cohort studies) or were the cases and controls (case-control studies) recruited over the same period of time?*

    For a study which does not specify the time period over which patients were recruited, the question should be answered as unable to determine.

Yes	1
No	0
Unable to determine	0

23. *Were study subjects randomised to intervention groups?*

    Studies which state that subjects were randomised should be answered yes except where method of randomisation would not ensure random allocation. For example alternate allocation would score no because it is predictable.

Yes	1
No	0
Unable to determine	0

24. *Was the randomised intervention assignment concealed from both patients and health care staff until recruitment was complete and irrevocable?*

    All non-randomised studies should be answered no. If assignment was concealed

from patients but not from staff, it should be answered no.

Yes	1
No	0
Unable to determine	0

25. *Was there adequate adjustment for confounding in the analyses from which the main findings were drawn?*

    This question should be answered no for trials if: the main conclusions of the study were based on analyses of treatment rather than intention to treat; the distribution of known confounders in the different treatment groups was not described; or the distribution of known confounders differed between the treatment groups but was not taken into account in the analyses. In nonrandomised studies if the effect of the main confounders was not investigated or confounding was demonstrated but no adjustment was made in the final analyses the question should be answered as no.

Yes	1
No	0
Unable to determine	0

26. *Were losses of patients to follow-up taken into account?*

    If the numbers of patients lost to follow-up are not reported, the question should be answered as unable to determine. If the proportion lost to follow-up was too small to affect the main findings, the question should be answered yes.

Yes	1
No	0
Unable to determine	0

Power

27. *Did the study have sufficient power to detect a therapist effect where the probability value for a difference being due to chance is less than 5%?*

How many therapists were there and how many patients did they treat?

Were there at least 10 therapists in total? Ideally the number of therapists should be maximised, with a minimum of 100 recommended, and at least 50 required for statistical significance. Did **all** therapists treat at least 10 patients?

≥100 therapists all treating ≥10 patients each	5
≥100 therapists with some or all treating treating <10 patients, or 50-99 therapists all treating ≥10 patients each	4
50-99 therapists with some or all therapists treating <10 patients	3
10-49 therapists all treating ≥10 patients	2
10-49 therapists with some or all therapists treating <10 patients	1
<10 therapists	0

Changes to original Downs & Black (1998) checklist:

- *6. Are the main findings of the study clearly described?*

  Changed from:
  Simple outcome data (including denominators and numerators) should be reported for all major findings so that the reader can check the major analyses and conclusions.
  (This question does not cover statistical tests which are considered below).

Yes	1
No	0

  Changed to:
  Simple outcome data should be reported for all major therapist effects so that the reader can check the major analyses and conclusions.
  (This question does not cover statistical tests which are considered below).

Yes	1
No	0

- *7. Does the study provide estimates of the random variability in the data for the main outcomes?*

  Changed from:
  In non normally distributed data the inter-quartile range of results should be reported. In normally distributed data the standard error, standard deviation or confidence intervals should be reported. If the distribution of the data is not described, it must be assumed that the estimates used were appropriate and the question should be answered yes.

Yes	1
No	0

  Changed to:
  In non-normally distributed data the inter-quartile range of results should be reported. In normally distributed data the standard error, standard deviation or confidence intervals should be reported around the therapist effect. If the distribution of the data is not described, it must be assumed that the estimates used were appropriate and the question should be answered yes.

Yes	1
No	0

- *18. Were the statistical tests used to assess the main outcomes appropriate?*

  Changed from:
  The statistical techniques used must be appropriate to the data. For example nonparametric methods should be used for small sample sizes. Where little statistical analysis has been undertaken but where there is no evidence of bias, the question should be answered yes. If the distribution of the data (normal or not) is not described it must be assumed that the estimates used were appropriate and the question should be answered yes.

Yes	1
No	0
Unable to determine	0

  Changed to (based on Baldwin & Imel, 2013):

  18. Were the statistical tests used to assess the therapist effects appropriate?

Were the data analysed within a hierarchical structure (e.g. using Multilevel Modelling), using random effects analysis, or at least involved calculation of the intraclass coefficient (ICC) for therapists?

Yes	1
No	0
Unable to determine	0

- *19. Was compliance with the intervention/s reliable?*

Changed from:
Where there was non compliance with the allocated treatment or where there was contamination of one group, the question should be answered no. For studies where the effect of any misclassification was likely to bias any association to the null, the question should be answered yes.

Yes	1
No	0
Unable to determine	0

Changed to:

19. Was compliance with the intervention/s assessed?

Where there was non-compliance with the allocated treatment or where there was contamination of one group, the question should be answered no. For studies where the effect of any misclassification was likely to bias any association to the null, the question should be answered yes.

Yes	1
No	0
Unable to determine	0

- *27. Did the study have sufficient power to detect a clinically important effect where the probability value for a difference being due to chance is less than 5%?*

Changed from:
Sample sizes have been calculated to detect a difference of x% and y%.

Changed to (based on Adelson & Owen, 2012; Baldwin & Imel, 2013; Hox, 2010 & Schiefele et al., 2017):

27. Did the study have sufficient power to detect a therapist effect where the probability value for a difference being due to chance is less than 5%?

How many therapists were there and how many patients did they treat?

Were there at least 10 therapists in total? Ideally the number of therapists should be maximised, with a minimum of 100 recommended, and at least 50 required for statistical significance. Did **all** therapists treat at least 10 patients?

≥100 therapists all treating ≥10 patients each	5
≥100 therapists with some or all treating treating <10 patients, or 50-99 therapists all treating ≥10 patients each	4
50-99 therapists with some or all therapists treating <10 patients	3
10-49 therapists all treating ≥10 patients	2
10-49 therapists with some or all therapists treating <10 patients	1
<10 therapists	0

Appendix B - Quality Checklist Results

Quality checklist results from main rater

Type of study	Author(s) and date	Question number																											
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	Total
RCT	Erickson et al. (2012)	1	1	1	1	1	1	0	0	1	1	U/D	U/D	1	0	0	1	1	1	1	1	1	1	1	0	1	1	1	20
	Goldsmith et al. (2015)	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	0	1	1	0	22
	Moyers et al. (2016)	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	24
	Owen et al. (2015)	1	1	0	1	1	1	1	0	1	0	1	1	1	0	0	1	1	1	1	1	U/D	1	1	0	1	1	4	23
Naturalistic	Ali et al. (2014)	1	1	1	1	2	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	0	1	1	3	26
	Chow et al. (2015)	1	1	1	0	2	1	0	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	4	24
	Firth et al. (2015)	1	1	1	1	2	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	4	26

Table B1 continued

Type of study	Author(s) and date	Question number																											
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	Total
Naturalistic	Goldberg et al. (2016b)	1	1	1	1	2	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	5	27
	Green et al. (2014)	1	1	1	1	2	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	1	23
	Hayes et al. (2015)	1	1	1	0	1	1	0	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	1	20
	Hayes et al. (2016)	1	1	1	0	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	4	24
	Kraus et al. (2016)	1	1	1	0	2	1	0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	4	23
	Laska et al. (2013)	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	1	22
	Nissen-Lie et al. (2016)	1	1	1	0	1	1	1	0	1	0	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	5	25
	Owen et al. (2016)	1	1	1	0	1	1	0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	4	22

Table B1 continued

Type of study	Author(s) and Date	Question number																										
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	26	Total
Naturalistic	Pereira et al. (2017)	1	1	1	1	2	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	2	24
	Saxon & Barkham (2012)	1	1	1	1	2	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	5	27
	Saxon et al. (2017)	1	1	1	1	2	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	4	26
	Schiefele et al. (2017)	1	1	1	1	2	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	4	26
	Wiborg et al. (2012)	1	1	1	1	2	1	1	0	1	0	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	1	22

Note. shaded area denotes less than maximum score. U/D = unable to determine; N/A = not applicable

Table B2

Quality checklist ratings – independent raters

	Question number																											
Author(s) and Date	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	Total
Rater 1																												
Erickson et al. (2012)	1	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1	1	1	1	1	1	1	0	1	0	1	19
Pereira et al. (2017)	1	1	1	1	2	1	1	0	1	1	1	1	0	0	0	1	1	1	1	1	0	1	0	0	0	0	2	20
Saxon & Barkham (2012)	1	1	1	0	2	1	1	0	1	0	1	1	1	0	0	1	1	1	1	1	0	1	0	0	0	0	5	22
Wiborg et al. (2012)	1	1	1	1	2	1	1	0	1	0	1	1	1	0	0	1	1	1	1	1	0	1	0	0	1	0	1	20
Rater 2																												
Saxon et al. (2017)	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	0	1	1	1	0	0	1	1	4	24
Hayes et al. (2015)	1	1	1	0	1	0	1	0	1	1	1	1	1	0	0	1	1	1	0	1	1	1	0	0	1	1	1	19
Goldsmith et al. (2015)	1	1	1	1	0	1	1	0	0	1	1	1	1	0	0	1	1	1	0	0	0	0	1	0	0	0	0	14
Laska et al. (2013)	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	0	1	1	1	0	0	1	1	1	21

Note. Both independent raters showed substantial agreement with the original rater using Cohen's kappa (κ) for inter-rater reliability. For rater 1, $\kappa=0.72$ and rater 2, $\kappa=0.66$ (both $p<0.01$).